M-QUEST - Meme Question-Understanding Evaluation on Semantics and Toxicity

Anonymous ACL submission

Abstract

Internet memes are a powerful form of online communication, yet their multimodal nature and reliance on commonsense and cultural knowledge make automated toxicity detection challenging. Prior work has studied individual dimensions of memes: textual content via optical character recognition, visual content via object recognition, and toxicity via hate speech detection, but lacks a unified architecture that formally identifies the semantic dimensions contributing to meme toxicity. We present three contributions: (1) a semantic framework identifying the dimensions necessary for meme interpretation; (2) a corresponding benchmark, M-QUEST, of 609 question-answer pairs for 307 memes, generated via a semi-automatic process grounded in the framework; and (3) an empirical evaluation of 12 VLMs on M-QUEST. The results show that the instruction-tuned models with reasoning capabilities significantly outperform others, whereas pragmatic inference remains challenging across all models.

Disclaimer: This paper contains discriminatory and explicit content that may be disturbing to some readers.

1 Introduction

Internet memes are "linguistic, image, audio, and video texts created, circulated, and transformed by countless cultural participants" (Milner, 2018). Memes are currently a key form of communication in online public and private conversations. From funny content, they become actual "modes of hypersignification, wherein the code itself becomes the focus of attention" (Shifman, 2014). The communicative power of memes resides in their ability to condense complex ideas into shareable formats (Onielfa et al., 2022). This power is sustained by two factors: (i) resilience over time due to their multimodal nature, and (ii) traction gained from cultural and social overproduction, amplified by users consuming multimodal content in binge-watching patterns on social media (Shifman, 2019). While memes usually express humor, many "harmful memes" (Pramanick et al., 2021a) or "hateful memes" (Qu et al., 2023), jointly termed "toxic" memes (Pandiani et al., 2025), are designed or used for malicious purposes. Their scale, velocity, and variety make identifying toxicity in memes challenging for content moderators, and inspires work on AI-assisted toxicity detection (Ilievski, 2024).

Early AI research treated meme toxicity as a black-box classification task (Kiela et al., 2020; Fersini et al., 2022). Yet, to paraphrase Tolstoy (1878): "all happy memes are alike, each toxic meme is toxic in its own way." Identifying toxicity features requires a conceptual decomposition that goes beyond the meme as a material object (image + text) to the meme as a cultural object: a relational entity connecting observers, their background knowledge, and the broader cultural context implicated in its interpretation (Bollici et al., 2026).

This recognition of the multi-faceted nature of meme toxicity has led to recent research at the intersection of toxicity and reasoning (Pandiani et al., 2025). Semantic frameworks that characterize memes have been proposed in the form of theoretical frameworks (Bollici et al., 2026) or as knowledge graphs (Tommasini et al., 2023; Martinez Pandiani et al., 2025). Meanwhile, benchmarks evaluate the ability of vision-language models (VLMs) to assess and explain toxicity, in relation to their targets (Pramanick et al., 2021b), entity roles (Sharma et al., 2022b), background knowledge (Bhowmick et al., 2021), and emotions (Alzu'bi et al., 2023). Other benchmarks aim at a more comprehensive set of features (Nguyen et al., 2025), or alignment with human values (Lin et al., 2024).

Two limitations of prior work emerge. First, as shown in Table 1, *evaluation dimensions across meme toxicity benchmarks remain discrepant and seemingly ad-hoc*, highlighting the need for com-

Table 1: **Meme dimensions have been sparsely addressed in prior work, lacking a comprehensive view**. We summarise the most relevant works overlapping with our evaluation framework and M-QUEST benchmark. See Appendix for the extended Table 4 with 40 analysed works.

Paper Reference	Toxicity	Textual Material	Visual Material	Scene	Background Knowledge	Overall Intent	Emotion	Target Community	Semiotic Projection	Analogical Mapping
(Hee et al., 2023)	✓	✓	✓	×	✓	✓	✓	✓	×	×
(Qu et al., 2022)	✓	✓	✓	✓	✓	✓	✓	×	×	✓
(Rajput et al., 2022)	✓	✓	✓	×	×	×	×	×	×	×
(Gasparini et al., 2022)	✓	✓	×	×	×	✓	×	✓	×	×
(Ramamoorthy et al., 2022)	×	✓	✓	✓	×	✓	✓	×	×	×
(Kiela et al., 2020)	✓	✓	✓	×	×	✓	×	✓	×	×
(Hee et al., 2023)	✓	✓	✓	×	✓	×	✓	✓	×	×
(Fersini et al., 2022)	✓	✓	✓	×	×	×	×	×	×	×
(Sharma et al., 2022a)	✓	✓	✓	✓	✓	×	×	✓	×	×
(Nguyen et al., 2025)	×	✓	✓	×	✓	✓	✓	✓	×	×
M-QUEST (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

prehensive frameworks grounded in theoretical conceptualizations (Bollici et al., 2026). Second, there is a *lack of separation between toxicity identification and reasoning*: analogously to work that separates perception from reasoning (Jiang et al., 2024), principled evaluation must isolate individual toxicity causes to expose the granular capabilities and limitations of state-of-the-art VLMs.

To that end, we provide *a comprehensive framework for connecting the toxicity of internet memes to underlying semantic dimensions and assessing the ability of VLMs to analyze and explain toxicity through these dimensions*. Our contribution is threefold: 1. A **framework** that formalizes and operationalizes a comprehensive set of 10 dimensions, providing a semantic map of the features in existing literature. 2. A compositional **benchmark** called M-QUEST with 609 paired multiple-choice questions about the toxicity and its underlying reasons of 307 memes. M-QUEST is built semi-automatically by combining our framework dimensions, the divergent generative abilities of VLMs, and the common sense of human judges. 3. An **empirical analysis** of the granular performance of 12 diverse open-weight VLMs on the 10 dimensions of M-QUEST.

2 Related Works

As apparent from Table 1, existing toxicity work has focused on only a few dimensions beyond Textual and Visual Material. Our framework subsumes all of these while introducing *Semiotic Projection* - identifying which meme element the observer or external entities are projected onto - as a novel contribution absent from prior research on memes. We present here benchmarks and approaches grounding our framework. More specific references per each dimension are given in Section 3.

Meme Benchmarks Several benchmarks address meme toxicity subtasks (Table 1).[1] The Hateful Memes challenge (Kiela et al., 2020) is a foundational and widely used benchmark to date. Gasparini et al. (2022) target misogyny specifically (800 memes), as does MAMI SemEval-2022 (Fersini et al., 2022), which adds misogyny subclassification. MOMENTA (Pramanick et al., 2021b) classifies meme harmfulness and its target. The GOAT-Benchmark (Lin et al., 2024) covers 6,000 memes across five hate speech types (hatefulness, misogyny, offensiveness, sarcasm, harmful content) via binary classification. The closest to our work is MemeQA (Nguyen et al., 2025), comprising $9,000$ human-curated multiple-choice Q&A across seven dimensions.

Our benchmark M-QUEST subsumes a broader set of dimensions than all prior work, including MemeQA. It provides a more granular analysis by explicitly representing relations among elements within each dimension and across dimensions with respect to toxicity. Finally, it is compositional: toxicity labels are systematically linked to explanation

[1] See (Pandiani et al., 2025) for a comprehensive survey.

questions drawn from the other nine dimensions.

3 Framework of Meme Dimensions

Table 1 consolidates the key semantic dimensions considered in prior work on meme toxicity and related tasks, highlighting that they have been considered in seemingly arbitrary subsets, lacking a comprehensive view. Based on this body of work, we design a framework including 10 dimensions. These dimensions represent cognitive elements employed in the sense-making process of understanding a meme and its toxicity. We now describe and motivate each dimension.

Toxicity Harmful, offensive, or problematic content in the meme. Following (Pandiani et al., 2025), we distinguish six subtypes: *Abuse* (threats or abuse toward individuals or communities), *Cyberbullying* (mocking based on protected characteristics), *Harmfulness* (potential to harm individuals, organizations, or society), *Hatefulness* (direct or indirect attacks based on protected characteristics), *Misogyny* (hate targeting women), and *Propaganda* (content designed to influence opinions or actions toward a goal). Our notion of toxicity is linked to - and grounded in - all other 9 dimensions.

Textual Material Any visible text in the meme. Virtually all meme analysis works incorporate this component, focusing on captions, overlay text, and speech bubbles, particularly those focused on hate speech detection (Rajput et al., 2022; Bhandari et al., 2023; Kiela et al., 2020). Our notion includes also diegetic text (any text that exists within the meme image).

Visual Material Any visual entity in the meme: people, objects, characters, graphics, and salient visual elements. Alongside Textual Material, it is the most prevalent dimension; several works analyze both jointly (Zhang et al., 2023; Sabat et al., 2019; Badour and Brown, 2021). Rather than treating memes as images with embedded text (Bhandari et al., 2023), our approach integrates both modalities in a unified Knowledge Base (KB).

Scene The meaningful interpretation of how visual entities relate to one another, encompassing spatial arrangements, positioning, and layout. It has been investigated to detect victims of hateful memes (Sharma et al., 2022a) and for general multimodal meme question answering (Agarwal et al., 2024). Our notion subsumes relational knowledge about interactions of visual entities, with the scope of making sense of visual clues (e.g. starting from identifying `person`, `spoon` and `kitchen`, we build in the KB the scene "person cooking" taking as roles the above mentioned visual entities).

Background Knowledge Any external knowledge required to interpret the meme that is not directly observable. Our notion subsumes public personas (Bhowmick et al., 2021), contextual information (Shang et al., 2021b), factual correctness (Jabiyev et al., 2021), real-world named entities (Sharma et al., 2022a) and social entities (Pramanick et al., 2021a). It expands over specific groups (Qu et al., 2023), while extending coverage to general stereotypes, events, concepts, beliefs, and implicit references (e.g., actor names, historical events, political parties, cultural phenomena, or fictional references).

Overall Intent The communicative purpose behind the meme. The *Overall Intent* determines *underlying reasons* (Hee et al., 2023) for which a meme is meant to be toxic, political, nonsensical, etc. Our notion thus incorporates DisInfoMeme's focus on intentional misinformation (Qu et al., 2022) and Agarwal et al. (2024)'s notion of intended effects, expanding over Nguyen et al. (2025), further encompassing the meme's creator intentions, message goals, and purpose indicators.

Emotion Emotion conveyed by the meme. *Sentiment analysis* and *emotion detection*, are observed in a considerable body of work analyzing memes, including the Memotion 1.0 and 2.0 datasets (Sharma et al., 2020; Ramamoorthy et al., 2022). They are often integrated as plain labels of the meme (Maity et al., 2022), or to capture the emotion conveyed by the meme's visual or textual component. Notably, the Misogyny Memes benchmark (Gasparini et al., 2022) incorporates emotions among the features they encompass. Beyond using emotion analysis to detect offensive intent (Kumari et al., 2023; Alzu'bi et al., 2023), we include neuro-physiological expressors: facial expressions, emotional gestures, and affective language.

Target Community The intended audience or community targeted by the meme. While prior work on hateful memes focuses on specific individuals targeted by the attack of the meme (Mathias et al., 2021), being it towards individuals and organizations (Jha et al., 2024), or communities (Pramanick et al., 2021a,b), our notion broadens the

concept to include community markers, demographic cues, cultural references, and audience signals.

Semiotic Projection A novel dimension capturing the projection of the meme observer onto a meme element (typically visual, textual, or scene). While prior work addresses narrative role projection - Hero, Villain, and Victim - onto meme elements (Sharma et al., 2022b; Bollici et al., 2026), the semiotic triangle formed by (1) a visual entity, (2) its role in the meme, and (3) the observer's projection onto it remains underexplored. For instance, in the Disaster Girl meme,[2] the text `Me` juxtaposed with the girl projects the observer onto an apparently innocent character (the girl being a child) whose smirk hints at complicity in the house on fire in the background. The resulting inference is a visual manifestation of *Schadenfreude*: pleasure derived from another entity's misfortune.

Analogical Mapping A mapping between two meme entities - typically visual and textual - following the pattern `X is Y`. Inspired by analogical thinking (Holyoak and Thagard, 1996), the Structure-Mapping Engine (Forbus et al., 2017), and cognitive metaphor theory (Lakoff and Johnson, 2024), we unify analogical and metaphorical mappings under a single dimension. For instance, the text `Me` juxtaposed with a perplexed child, maps `Me` onto `child being perplexed`, yielding *I am perplexed*.[3] Spatial proximity commonly signals such mappings. A notable work investigating this dimension is Met-meme (Xu et al., 2022), a multimodal dataset rich in metaphorical features. More principled theoretical proposals (Bollici et al., 2026) and ontologies (Tommasini et al., 2023; Martinez Pandiani et al., 2025) are emerging, although to date they have not been operationalized into practical benchmarks or methods, and the explicit focus on mapping structure and semantic gain is a novelty of this work.

4 M-QUEST Benchmark Construction

In this section, we describe our semi-automatic benchmark construction method, starting with automatic Q&A generation, followed by a human validation. Finally, we describe the resulting benchmark statistics.

[2] https://knowyourmeme.com/memes/disaster-girl

[3] https://en.wikipedia.org/wiki/Side_Eyeing_Chloe

4.1 Automatic Q&A Generation

Automatic benchmark creation consists of two steps shown in Fig. 1: Knowledge Extraction and Q&A Generation.

Knowledge Extraction (KE) The entire framework is modular and component-agnostic. We process one meme at a time, extracting knowledge per dimension using our Dimensions Framework (§3) coupled with the Qwen3-VL-8b-Instruct VLM. Each VLM call receives the dimension definition integrated in the prompt and the meme as a `.png` file, yielding one knowledge base entry per dimension. Examples are shown in Fig. 1 (KE Snippet).

Prompts follow a fixed order, as in Table 5, Appendix A.3: basic dimensions first (Textual and Visual Material), then dependent ones (Scene), and finally those with complex inter-dimensional relations (Analogical Mapping, Semiotic Projection). At each step, previously extracted knowledge is included in the prompt, ensuring consistency and shared anchoring points across dimensions. The output is a comprehensive KB about the meme and its toxicity, according to the dimensions described in Section 3. A worked example is provided in Appendix A.3.

Candidate Q&A Generation Using the extracted knowledge, the VLM generates N questions per dimension, each with four answers: one correct, one plausible distractor, one implausible distractor, and a fixed "None of the others" option. Answers are shuffled to avoid positional bias, and a brief explanation accompanies each correct answer. Inputs to the VLM are the dimension definition, the previously extracted knowledge, and the meme image (Fig. 1).

Outcome We randomly selected 1,000 memes from the Hateful Memes dataset (Kiela et al., 2020). This dataset was selected for: 1. being widely adopted in previous harmful memes research, 2. including memes from different domains (e.g., Covid pandemics, elections) and 3. covering multiple types of toxicity (e.g., harmful, misogyny, etc.). Our KE step generates one KB per meme (statistics in Fig. 3). Creating one question per dimension yields 10,000 candidate Q&As in total (1 question $\times$ 10 dimensions $\times$ 1,000 memes).

4.2 Manual Data Validation

Human Validation Task Design Since our primary focus is toxicity, human judges validated the

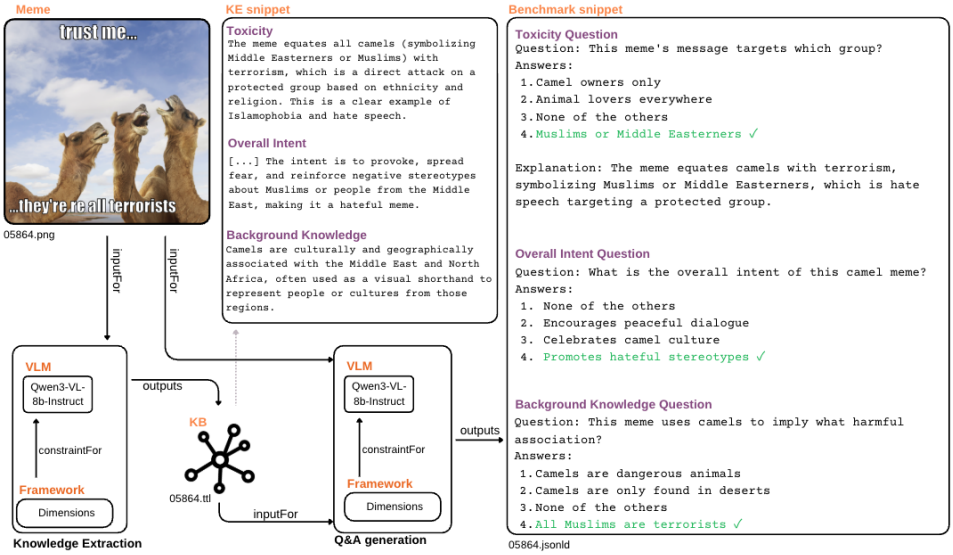


Figure 1: M-QUEST framework: a meme is given as input to a VLM, which generates a Knowledge Base (KB), used by a second VLM module to generate Q&A sets based on KB. Running example details in Appendix A.3.

Table 2: Q&As distribution across dimensions for different benchmark subsets. For the benchmark Extension, based on the original 307 memes, the number 180 in the "Unique Memes" column, should be interpreted as a subset of the 307 memes in the ToxM-QUEST row.

Subset	Toxicity	Textual Material	Visual Material	Scene	Background Knowledge	Overall Intent	Emotion	Analogical Mapping	Target Community	Semiotic Projection	Unique Memes	Avg. Q&As per Meme
ToxM-QUEST	307	10	14	9	9	9	—	—	—	—	307	1.16
Extension	—	35	35	35	26	26	30	28	20	16	180	1.48
M-QUEST	**307**	**45**	**49**	**44**	**35**	**35**	**30**	**28**	**20**	**16**	**307**	**1.98**

KBs and Q&A pairs for this dimension via a custom interface (Fig. 7). Per meme, judges evaluated up to six fields: (1) plausibility of the toxicity assessment; (2) if a toxicity-linked KB existed, plausibility of the associated element's assessment; (3–4) whether the toxicity-derived question was on point and its correct answer plausible; (5–6) the same two judgments for the question derived from the toxicity-linked element. This results in a minimum of one annotation per meme (toxicity question plausibility) and a maximum of six.

Outcome Fourteen annotators (9 male, 5 female), all familiar with LLMs, KBs, and memes, validated 1,000 memes in total focusing on the Toxicity dimension. In 307 cases, both the question and correct answer were positively validated, forming an intermediate dataset called ToxM-QUEST. To cover all dimensions, we sampled 50 additional questions per dimension from the 307 ToxM-QUEST memes and subjected them to human validation: 59.33% of memes received positive validation for both questions. Approval rates were highest for fundamental dimensions – Scene, Textual, and Visual Material (74–76%) – and lowest for Semiotic Projection and Target Community (32% and 40%, respectively). Further details are provided in Appendix A.4.

4.3 M-QUEST

Key Statistics Table 2 summarizes the resulting benchmark M-QUEST, which comprises 609 Q&A pairs for 307 memes – nearly two questions per meme: one toxicity question is always present, while another reasoning question depends on the connection between the reasoning question and toxicity one. The reasoning questions are relatively balanced across the nine dimensions, ranging from 16 (Semiotic Projection) to 35 (Visual Material, Textual Material, Scene).

Release and Licensing Following Meta's license on the Hateful Memes dataset, the M-QUEST repository includes generated graphs, human annotations, and Q&A pairs for 1,000 memes, but not the images themselves. The images are available from the Hateful Memes challenge repository,[4]

[4] https://ai.meta.com/tools/hatefulmemes/

with alignment ensured via original IDs.

5 Evaluation

Next, we conduct benchmarking experiments to gauge the performance of current state-of-the-art (SoTA) VLMs on M-QUEST. We present our experimental setup, followed by the obtained experimental results and a qualitative analysis.

5.1 Experimental Setup

Evaluation Metrics We report Q&A *accuracy* (percentage of correctly answered questions), both overall and per dimension. For memes with both a toxicity and a reasoning question, we compute *group accuracy* – whether a model answers both correctly, i.e., whether model answers correctly about toxicity for the right reasons. Performance across dimensions is summarized via *macro-average* (unweighted mean). This setup enables both overall comparison and fine-grained identification of each VLM's strengths and weaknesses.

Models We evaluate 12 open-source VLMs spanning four capability profiles: *Perception-focused* (no instruction tuning or advanced reasoning): BLIP2-Flan-T5-xl (Li et al., 2023) and Pixtral-12B (Agrawal et al., 2024); *Instruction-tuned*: InstructBLIP-Vicuna-7B (Dai et al., 2023), FastVLM-1.5B/7B (Vasu et al., 2025) and DeepSeek-VL-1.3B/7B-Chat (Lu et al., 2024); *Reasoning-oriented*: LLaVA-v1.5 (Liu et al., 2024a) and LLaVA-v1.6-Vicuna (Liu et al., 2024b); *Instruction-tuned with reasoning*: Qwen2-VL-7B (Wang et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025b), and Qwen3-VL-8B-Instruct (Bai et al., 2025a).

Model Outputs Model outputs are text strings (Fig. 6); valid answers are `A`, `B`, `C`, or `D`. Models capable of chain-of-thought reasoning (Qwen family, LLaVA family, Pixtral-12B) are prompted to justify their choice; others return only the final answer. Invalid responses trigger re-prompting until a valid option is returned, ensuring consistency across all models.

Implementation Details All models are evaluated zero-shot with greedy decoding ($temperature$=0, top-p=1.0). For multiple-choice questions, the answer choice is generated first (stopping upon a valid option), followed by the CoT explanation (terminated naturally by the end-of-sequence token). All experiments were run on a single H100 GPU, taking approximately 10 hours in total.

Table 3: Accuracy of VLMs on M-QUEST: accuracy for All questions, grouped per meme (T+R), Toxicity accuracy, and combined Reasoning accuracy across other dimensions, along with the Macro-Average. Models are grouped by capability profile.

VLM	All	Group (T+R)	T only	R only	Macro Avg.
Perception-focused					
BLIP2-Flan-T5-xl	13.2	12.3	14.7	14.7	14.3
Pixtral-12B	70.3	60.6	72.0	54.7	57.4
Instruction-tuned					
InstructBLIP-Vicuna-7B	22.9	18.7	22.8	17.8	20.6
FastVLM 1.5B	32.5	14.0	31.1	37.7	34.3
FastVLM 7B	44.0	19.9	35.1	53.0	48.6
DeepSeek-VL-1.3B	71.1	52.1	65.2	77.2	74.0
DeepSeek-VL-7B	81.3	67.4	78.2	84.4	82.2
Reasoning-oriented					
LLaVA-v1.6-Vicuna	47.3	46.8	50.2	40.2	41.5
LLaVA-v1.5	56.8	53.4	54.7	70.4	68.3
Instruction-tuned + reasoning					
Qwen3-VL-8B	83.8	78.3	80.1	**95.3**	**94.3**
Qwen2-VL-7B	85.6	82.3	84.4	94.7	93.8
Qwen2.5-VL-7B	**86.4**	**83.3**	**86.6**	94.7	93.3

5.2 Results

Performance Analysis As shown in Table 3, accuracy on M-QUEST ranges from below chance to 86%, confirming the benchmark's difficulty. Reasoning dimensions consistently score higher than toxicity-related ones, suggesting that explicit visual or textual reasoning, scene understanding, and background knowledge interpretation are easier than inferring latent toxic intent. Notably, Qwen models approach near-perfect accuracy on reasoning questions when toxicity is excluded. Still, performance drops when implicit inference is required as shown in Fig. 2, underscoring that *latent semantic understanding across modalities is systematically harder than surface-level cue matching*.

Substantial differences emerge across model families. BLIP2-Flan-T5-xl and InstructBLIP-Vicuna-7B fall below random chance, reflecting limited instruction tuning. FastVLM-1.5B and 7B achieve moderate scores, capturing basic visual features but lacking the multimodal alignment needed for relational reasoning. Within the LLaVA family, v1.5 outperforms v1.6-Vicuna, showing that instruction tuning differences matter even when the vision encoder family is similar. Pixtral-12B benefits from a CLIP encoder's semantically rich representations. DeepSeek-VL-1.3B and 7B outperform earlier instruction-tuned models, attributed to

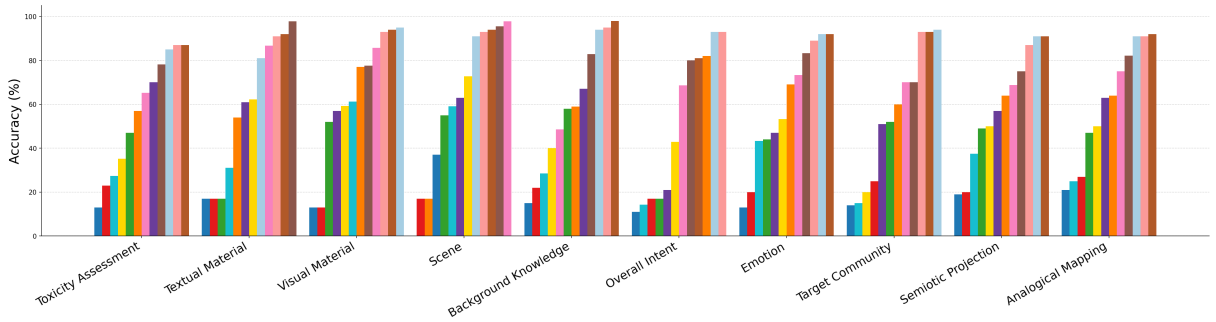


Figure 2: Accuracy of VLMs across dimensions, with BLIP2-Flan-T5-xl, InstructBLIP-Vicuna-7B, FastVLM-7B, FastVLM-1.5B, LLaVA-v1.6-Vicuna, LLaVA-v1.5, Pixtral-12B, DeepSeek-VL-1.3B-chat, DeepSeek-VL-7B-chat, Qwen3-VL-8B-Instruct, Qwen2-VL-7B-Instruct, Qwen2.5-VL-7B-Instruct.

stronger vision–language alignment. Interestingly, Qwen2.5-VL-7B slightly surpasses Qwen3-VL-8B, suggesting that model size alone is not decisive; rather, *models optimized for complex multi-step reasoning may occasionally overreason, leading to slightly lower accuracy in this setting*.

These trends are explained by two key factors. *Instruction tuning* enables models to align textual cues with visual information, while *reasoning capability* supports multi-step inference over contextual relationships. Models lacking both (BLIP2) struggle the most; instruction tuning alone (InstructBLIP) yields only moderate gains; and models combining both (Qwen family) achieve the strongest overall performance. This suggests that while reasoning capability is essential for interpreting complex meme semantics, excessive reasoning may introduce inefficiencies when tasks require only shallow inference.

Performance Across Evaluation Dimensions Group accuracy (correctly answering both the Reasoning and Toxicity questions for the same meme) reflects cross-dimensional consistency rather than isolated task performance. As shown in Table 3, group accuracy is generally lower than reasoning and toxicity accuracy, indicating that *aligning Reasoning outcomes with Toxicity judgments remains challenging*. Weak models (BLIP2, InstructBLIP) show low group accuracy due to limited multimodal alignment. Stronger models achieve higher group accuracy, suggesting improved cross-dimensional consistency when reasoning about the meme and evaluating its toxicity within the same context.

Per-dimension analysis (Fig. 2) reveals consistent patterns. Qwen models achieve near-perfect accuracy on perceptual dimensions (Scene, Visual Material) but weaker performance on latent relational dimensions (Toxicity, Overall Intent). LLaVA-v1.5 follows a similar pattern. BLIP2 and InstructBLIP represent the lower bound across most dimensions, failing to capture implicit vision–text relationships. Notably, Qwen-family models achieve near-perfect accuracy on Background Knowledge, suggesting that well-aligned instruction tuning enables effective external knowledge retrieval. Overall, *accurate perceptual recognition alone does not guarantee strong performance on latent relational dimensions; effective instruction tuning and multimodal reasoning remain essential for robust pragmatic inference*.

Implications The results highlight four key implications. First, *multimodal alignment and reasoning quality matter more than model size*. Second, models with stronger multimodal representations achieve consistently higher performance, highlighting the *importance of semantically rich visual–textual representations* for socially complex content: improved semantic grounding could substantially benefit toxicity analysis. Third, the gains from combining instruction tuning with reasoning training confirm that *toxic meme understanding is reasoning-intensive*: accurately inferring toxic intent requires integrating visual, textual, and contextual cues, which models lacking reasoning capability fail to capture. Finally, while current VLMs are approaching saturation on surface-level visual dimensions, *latent intent and toxicity remain challenging*: a critical gap for real-world deployment, where subtle harmful content remains most likely to be missed.

5.3 Qualitative Analysis

VLM Comparison on Toxicity Reasoning Considering the meme in Fig. 1, Box 1 presents a case study comparing four CoT-capable models with increasing performance on the same meme. InstructBLIP correctly flags hateful group dynamics but misses the symbolic camel–Muslim association, treating the text as directly describing the visual content. LLaVA-1.6 reaches the same conclusion, again reducing the meme to "camels as terrorists" without deeper inference. Pixtral-12B draws the correct connection between the visual entity (camel) and the target group (Muslims) via background knowledge, and correctly identifies the meme as perpetuating a harmful stereotype. Qwen3 provides the most nuanced interpretation, contextualizing "terrorist" both semantically and culturally, and building a layered, consistent reading that correctly identifies the adoption of a harmful stereotype.

Box 1 – Q: This meme's message targets which group?

InstructBLIP-Vicuna-7B CoT

The meme targets the group of people who are being accused of being terrorists, which is the group of Muslims or Middle Easterners. The meme is making a joke by saying "trust me, they're not terrorists" and is poking fun at the stereotype that all camels are terrorists.

LLava-1.6-Vicuna CoT

The text above the camels reads "trust me... they're all terrorists" which is a statement that is often used to stereotype and demonize Muslims or Middle Easterners. The humor in the meme is derived from the absurdity of camels being terrorists, which is a play on the idea that all Muslims are terrorists. The meme's message targets a specific group, which is Muslims or Middle Easterners, as it uses a stereotype to make a joke. The humor is derived from the absurdity of the situation, but it is important to recognize that such stereotypes are harmful and perpetuate negative attitudes towards a specific group of people.

Pixtral-12B CoT ✓

The meme uses a stereotype about camels being associated with the Middle East or desert regions. The phrase "they're all terrorists" is a generalization that implies all members of a certain group are terrorists, which is a derogatory and offensive statement. The context and the visual elements (camels in a desert) strongly suggest that the meme is targeting individuals from the Middle East or Muslims.

Qwen3-VL-8B-Instruct CoT ✓

The word "terrorists" is a highly charged term historically and politically associated with acts of violence, particularly those committed by individuals or groups from specific regions or religions; most notably, in Western discourse, often linked to Muslims or Middle Easterners. The meme uses camels, animals often culturally associated with the Middle East, as the speakers' of this harmful generalization. [...] The target of the meme's message is therefore the group that the word "terrorists" is being applied to in a sweeping, unjust manner, which, in this context, is Muslims or Middle Easterners, as that is the most common and historically charged association with the term in global discourse.

VLM Behavior Across Dimensions Fig. 9 illustrates how Qwen 2/2.5/3 reason differently across dimensions on the same meme. On the Scene dimension, all models correctly identify the salient visual elements (individuals in fruit costumes, people walking along a street) and infer the depicted activities. On the Toxicity dimension, however, all models incorrectly interpret the meme as implying fruit scarcity in Africa, missing its classification as false equivalence: the meme conflates dietary patterns with quality of life, reinforcing colonialist stereotypes.

These results reveal a recurring vulnerability: models performing adequately on unimodal or perceptual dimensions frequently fail when multimodal cue synthesis is required for toxicity assessment. LLaVA and Qwen3 show stronger recognition of implicit social critique and ironic framing, while earlier Qwen iterations exhibit a literal interpretation bias. These findings underscore the need for multi-dimensional evaluation to comprehensively assess VLMs' capacity for nuanced multimodal reasoning.

6 Conclusions

This paper studied the ability of state-of-the-art open-weight VLMs to assess toxicity and support their decisions by principled reasoning. To this end, we designed a semantic framework that consolidates 10 dimensions from prior work, including the understudied dimensions of analogical mapping and semiotic projection. Then, we designed a compositional benchmark, M-QUEST, with 307 memes and 609 questions, focusing on toxicity and its grounding in the interpretation of a second key dimension for each meme. Our experiments with 12 VLMs revealed substantial and systematic differences in the model performance. Models that incorporate instruction-tuning and reasoning significantly outperformed the others, especially on inference dimensions. This analysis revealed that meme toxicity assessment is an inherently knowledge-and reasoning-intensive task that requires rich multimodal representations, intent modeling, and commonsense reasoning. A key future direction is to extend the benchmark with dimensions that can further capture nuance in annotator judgments, namely, more granular background knowledge (Ilievski, 2024) and considerations of morality, culture, and legality (Forbes et al., 2020).

7 Limitations

Benchmark Size and Languages M-QUEST comprises 609 questions in total (307 on toxicity, 16–49 per other semantic dimension). Its limited size may constrain the generalisability of the findings. Similarly, its findings are based on English memes and may not fully extend to other languages. Expanding both the scale and linguistic diversity of M-QUEST is a key future direction.

Annotator Effects Human validation is inherently influenced by annotators' values, culture, and background. Our 14 judges are LLM-familiar scientists primarily of European origin, whose approval decisions may not generalize to other populations. Validation rates (Fig. 5) are further affected by cultural misalignment with the original memes, many of which reference US politics and everyday situations. Characterizing annotator (dis)agreement and its sources remains an important direction for future work.

Framework Design Our framework is built bottom-up by surveying existing semantic frameworks for memes. Consequently, dimensions such as target metadata (Tommasini et al., 2023), template configurations (Bates et al., 2025), and considerations of morality, ethics, culture, and legality (Forbes et al., 2020) are not included but warrant inclusion in future extensions. Alternative design approaches (e.g., combining top-down theoretical frameworks with bottom-up benchmark analysis) remain an important direction for future work.

8 Ethical Considerations

This work involves the analysis and automated processing of toxic internet memes, including content that may be harmful, offensive, or distressing. We acknowledge several ethical considerations. We further added a disclaimer after the Abstract.

Dataset and Content M-QUESTis derived from the Hateful Memes dataset (Kiela et al., 2020), which contains content that is intentionally hateful, offensive, or toxic. We do not redistribute the meme images, requiring users to obtain them directly from the original source in accordance with Meta's license. Researchers accessing M-QUESTshould be aware of the potentially distressing nature of the content.

Annotator Welfare Our 14 human annotators were exposed to toxic and hateful content as part of the validation process. Annotators were informed of the nature of the task prior to participation and provided consent. We recommend that future annotation efforts involving similar content implement structured exposure limits and access to psychological support.

Annotator Bias As noted in Section 7, our annotators are primarily of European origin with familiarity with LLMs and memes. Their judgments may not be representative of broader or more diverse populations, and cultural misalignment with the predominantly American cultural references in the dataset may have influenced validation rates.

Dual-Use Risk While our framework is designed to advance toxicity detection, the dimensions and knowledge extraction pipeline could in principle be repurposed to assist in the *creation* of more effective toxic memes. We encourage responsible use of this framework strictly for detection, moderation, and research purposes.

Model Evaluation The VLMs evaluated in this work may themselves produce harmful or biased outputs when processing toxic content. Our evaluation is conducted in a controlled research setting and is not intended to endorse or promote any specific model for deployment in content moderation systems without further safety evaluation.

Scope of Toxicity Our operationalization of toxicity follows (Pandiani et al., 2025) and is inherently partial. Definitions of harmful content are culturally situated and may not generalise across languages, communities, or legal jurisdictions. M-QUESTshould not be treated as a definitive or universal standard for toxicity classification.

Use of Language Models Claude Sonnet 4.6 was used solely for English proofreading and language checking of this paper. All VLMs mentioned in Section 5.1 were used exclusively for the benchmark evaluation tasks described in the paper, and no generative model was used to produce the scientific content of this work.

9 Responsible NLP Research Checklist

A1. Did you describe the limitations of your work? Yes, in the dedicated Section 7 - Limitations.

A2. Did you discuss any potential risks of your work? Yes, in dedicated Section 8 - Ethical Concerns.

B1. Did you cite the creators of artifacts you used? We created a benchmark, which is provided as additional files. We do not disclose the name of the creators in order not to violate anonymity.

We reused the Hateful Memes existing dataset, which is properly cited in the paper, and, according to their material license, only a limited subset (max 100 images) of their data can be released publicly. We fully comply with their license.

B2. Did you discuss the license or terms for use and / or distribution of any artifacts? We discussed among us and we'll be release it under CC-BY 4.0 license when anonymity will be dropped, after the review period.

B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)? Yes, our use of existing artifacts is consistent.

Yes, we specify intended use and it is fully compatible with original resources.

B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it? This point does not apply to our work.

B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.? Yes, in Section 2, and Section 4.3.

B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Yes, in Section 4.3, Section 5.1, and Section 5.3. Additionally further insights are provided in Appendix.

C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used? Yes, in Section 5.1.

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values? Yes, in Section 5.1.

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run? Yes, in Section 5.1, Section 5.3, and additionally in Appendix.

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, Spacy, ROUGE, etc.), did you report the implementation, model, and parameter settings used? Yes, in Section 5.1 and Section 3.

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.? Yes, it is described in Section 5.1, and screenshots are provided in Appendix.

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)? Yes, annotators were volunteers, and we specify gender and collective provenance, maintaining anonymity.

D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? This point does not apply to our work.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board? This point does not apply to our work.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data? Yes, in Section 4.3.

E1. If you used any AI assistants, did you include information about your use? Yes, in Section 8.

References

Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. 2024. Mememqa: multimodal question answering for memes via rationale-based inferencing. *arXiv preprint arXiv:2405.11215*.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Ahmad Alzu'bi, Lojin Bani Younis, Abdelrahman Abuarqoub, and Mohammad Hammoudeh. 2023. Multimodal deep learning with discriminant descriptors for offensive memes detection. *ACM Journal of Data and Information Quality*, 15(3):1–16.

Jamshid Bacha, Farman Ullah, Jebran Khan, Abdul Wasay Sardar, and Sungchang Lee. 2023. A deep learning-based framework for offensive text detection in unstructured data for heterogeneous social media. *IEEE Access*, 11:124484–124498.

Jafar Badour and Joseph Alexander Brown. 2021. Hateful memes classification using machine learning. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8. IEEE.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Luke Bates, Peter Ebert Christensen, Preslav Nakov, and Iryna Gurevych. 2025. A template is all you meme. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10443–10475.

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1994–2003.

Rajat Subhra Bhowmick, Isha Ganguli, Jayanta Paul, and Jaya Sil. 2021. A multimodal deep framework for derogatory social media post identification of a recognized person. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–19.

Gabriella Bollici, Filip Ilievski, and Wouter van Atteveldt. 2026. Political memes as fragmented narratives: An interdisciplinary framework for meme analysis. In *The Annual International Communication Association (ICA) Conference*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Mithun Das and Animesh Mukherjee. 2023. Banglaabusememe: A dataset for bengali abusive meme classification. *arXiv preprint arXiv:2310.11748*.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. *arXiv preprint arXiv:2109.08013*.

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.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

Kenneth D Forbus, Ronald W Ferguson, Andrew Lovett, and Dedre Gentner. 2017. Extending sme to handle large-scale cognitive modeling. *Cognitive Science*, 41(5):1152–1201.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*, 44:108526.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the underlying meaning of multimodal hateful memes. *arXiv preprint arXiv:2305.17678*.

Keith J Holyoak and Paul Thagard. 1996. *Mental leaps: Analogy in creative thought*. MIT press.

Filip Ilievski. 2024. *Human-Centric AI with Common Sense*. Springer.

Bahruz Jabiyev, Jeremiah Onaolapo, Gianluca Stringhini, and Engin Kirda. 2021. e-game of fame: Automatic detection of fake memes. In *TTO*, pages 1–11.

Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. 2024. Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations. *arXiv preprint arXiv:2401.09899*.

Yifan Jiang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, Jay Pujara, and 1 others. 2024. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. *Advances in Neural Information Processing Systems*, 37:46567–46592.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2023. Emoffmeme: identifying offensive memes by leveraging underlying emotions. *Multimedia Tools and Applications*, 82(29):45061–45096.

Gitanjali Kumari, Anubhav Sinha, Asif Ekbal, Arindam Chatterjee, and Vinutha B N. 2024. Enhancing the fairness of offensive memes detection models by mitigating unintended political bias. *Journal of Intelligent Information Systems*, 62(3):735–763.

George Lakoff and Mark Johnson. 2024. *Metaphors we live by*. University of Chicago press.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2024. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *ACM Transactions on Intelligent Systems and Technology*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

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

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1739–1749.

Delfina S Martinez Pandiani, Erik Tjong Kim Sang, and Davide Ceolin. 2025. Ontoxkg: An ontology-based knowledge graph of toxic symbols and their manifestations. In *International Conference on Web Engineering*, pages 119–127. Springer.

Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Talat. 2021. Findings of the woah 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 201–206.

Ryan M Milner. 2018. *The world made meme: Public conversations and participatory media*. mit Press.

Karthick Myilvahanan, B Shashank, Tushar Raj, Chiraanth Attanti, and Shivam Sahay. 2023. A study on deep learning based classification and identification of offensive memes. In *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pages 1552–1556. IEEE.

Khoi PN Nguyen, Terrence Li, Derek Lou Zhou, Gabriel Xiong, Pranav Balu, Nandhan Alahari, Alan Huang, Tanush Chauhan, Harshavardhan Bala, Emre Guzelordu, and 1 others. 2025. Memeqa: Holistic evaluation for meme understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18926–18946.

Thanh Tin Nguyen, Nhat Truong Pham, Ngoc Duy Nguyen, Hai Nguyen, Long H Nguyen, and Yong-Guk Kim. 2022. Hcilab at memotion 2.0 2022: Analysis of sentiment, emotion and intensity of emotion classes from meme images using single and multi modalities (short paper). In *DE-FACTIFY@ AAAI*.

Carles Onielfa, Carles Casacuberta, and Sergio Escalera. 2022. Influence in social networks through visual analysis of image memes. In *Artificial Intelligence Research and Development*, pages 71–80. IOS Press.

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

Kim Ngan Phan, Guee-Sang Lee, Hyung-Jeong Yang, and Soo-Hyung Kim. 2022. Little flower at memotion 2.0 2022: Ensemble of multi-modal model using

attention mechanism in memotion analysis. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*.

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

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. 2022. Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation. *arXiv preprint arXiv:2205.12617*.

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

Kshitij Rajput, Raghav Kapoor, Kaushal Rai, and Preeti Kaur. 2022. Hate me not: detecting hate inducing memes in code switched languages. *arXiv preprint arXiv:2204.11356*.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, and 1 others. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *Proceedings of De-Factify: workshop on multimodal fact checking and hate speech detection, CEUR*, volume 17.

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv preprint arXiv:1910.02334*.

Lanyu Shang, Christina Youn, Yuheng Zha, Yang Zhang, and Dong Wang. 2021a. Knowmeme: A knowledge-enriched graph neural network solution to offensive meme detection. In *2021 IEEE 17th International Conference on eScience (eScience)*, pages 186–195. IEEE.

Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. 2021b. Aomd: An analogy-aware approach to offensive meme detection on social media. *Information Processing & Management*, 58(5):102664.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020. Semeval-2020 task 8: Memotion analysis–the visuolingual metaphor! *arXiv preprint arXiv:2008.03781*.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022a. Disarm: Detecting the victims targeted by harmful memes. *arXiv preprint arXiv:2205.05738*.

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

Limor Shifman. 2014. The cultural logic of photo-based meme genres. *Journal of visual culture*, 13(3):340–358.

Limor Shifman. 2019. Internet memes and the twofold articulation of values. *Society and the internet: How networks of information and communication are changing our lives*, pages 43–57.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020a. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2023. Trollswithopinion: A taxonomy and dataset for predicting domain-specific opinion manipulation in troll memes. *Multimedia Tools and Applications*, 82(6):9137–9171.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020b. A dataset for troll classification of tamilmemes. In *Proceedings of the WILDRE5–5th workshop on indian language data: resources and evaluation*, pages 7–13.

Leo Tolstoy. 1878. *Anna karenina*. The Russian Messenger. First published in Russian 1873–1877.

Riccardo Tommasini, Filip Ilievski, and Thilini Wijesiriwardene. 2023. Imkg: The internet meme knowledge graph. In *European Semantic Web Conference*, pages 354–371. Springer.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and 1 others. 2025. Fastvlm: Efficient vision encoding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19769–19780.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. 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*, pages 2887–2899.

Ning Zhang, Feng Xuan, Gu Tianlong, and Chang Liang. 2023. Mvlp: Multi-perspective vision-language pre-training model for ethically aligned meme detection. *TechRxiv arXiv:10.36227/techrxiv.23689437.v1*.

A Appendix

A.1 Overview of Toxicity Benchmarks

Table 4 provides an overview of toxicity work through the lens of the dimensions in our framework. Note that all of these dimensions, except for the last one, are covered by our methodological framework. The final dimension is template structure, The *Template* feature is focused on representing the meme image's structural organization. Some example of Template patterns are: N-panels, vertical or horizontal progression, layout templates for `X vs Y`, dialogical organization, and structural patterns. Several previous works focus on multi-faceted notions of meme templates (Sharma et al., 2022a; Mathias et al., 2021; Pramanick et al., 2021a,b). This dimension is not covered in the current version of our framework; we detail possible related research paths in the Conclusions and Future Works Section.

A.2 Additional Related Work: Theoretical Frameworks Beyond Toxicity

Other research beyond toxicity benchmarks has focused on capturing the semantics of internet memes. Here, the Internet Meme Knowledge Graph (IMKG) (Tommasini et al., 2023) models textual, visual, and external knowledge about memes by connecting three sources. A key resource in IMKG is KnowYourMeme,[5] which catalogs internet meme templates and instances together with background knowledge on their emergence. OnToxKG (Martinez Pandiani et al., 2025) contributes an ontology-grounded knowledge graph of nearly 800 symbols, and shows its impact on downstream analysis of toxicity. Bollici et al. (2026) operationalise the notion of narratives in memes into 1) entities, roles, and social identities; 2) legitimation strategies; and 3) emerging narratives. Pandiani et al. (2025) organize prior work on toxicity in memes, yielding a taxonomy of toxicity categories, and defining three core dimensions of toxicity: intent (why), tactic (how), and target (who). Our semantic framework and associated benchmark M-QUESTbuilds on these theoretical insights to provide a practical tool for characterizing hate speech via underlying semantic dimensions and for evaluating the ability of VLMs to discover such links automatically.

[5] https://knowyourmeme.com/

A.3 Running Example for our Extraction Procedure

In the running example shown in Fig. 1, meme 05864.png from the Hateful Memes dataset depicts three camels on in a desert-like scenario, with upper text: `trust me...`, followed by lower text: `...they're all terrorists`.

The meme is passed to the KE module, where Qwen3-VL-8b-Instruct is prompted to extract the dimensions sequentially. The prompt order is shown in Table 5: Textual and visual dimensions are extracted first, and the knowledge material is stored and passed in the prompts for subsequent dimensions. The output is a knowledge base. A snippet of the generated Knowledge Base is shown in Fig. 1. Among others, three dimensions are displayed: Toxicity, Overall Intent, and Background Knowledge. The Toxicity is described as such: *The meme equates all camels (symbolizing Middle Easterners or Muslims) with terrorism, which is a direct attack on a protected group based on ethnicity and religion. This is a clear example of Islamophobia and hate speech.*

Further knowledge extracted for Overall Intent and Background Knowledge dimensions is shown below Toxicity in Fig. 1.

The knowledge base is stored, and salient knowledge material is passed to the Q&A Generation component. For the Toxicity dimension, the above mentioned chunk of text is passed, successfully resulting in the question: `This meme's message targets which group?`, with the correct answer ($a4$: `Muslims or Middle Easterners`) and distractors ($a1$: `Camel owners only`; $a2$: `Animal lovers everywhere`; and $a3$: `None of the others`). Finally, the provided explanation justifies answer $a4$ as correct, identifying a mapping between camels and the term "terrorists," with background knowledge reference to the stereotype of Muslims being terrorists (possibly following 9/11 terrorist attack).

Modularity and Component Agnosticism Note that the entire framework is modular and component-agnostic: the knowledge base can be enriched with additional modules, such as refined modules focused on emotions, adopting a theory-driven approach and reusing well-established emotion theories as in (Maity et al., 2022). The more refined dimension definitions would then be used to prompt the VLM. The neural component—in this case a VLM—could be replaced with other spe-

Table 4: **Memes dimensions have been sparsely and non-systematically analysed**. Here the full table of the 40 works grounding our framework used to build M-QUEST benchmark.

Paper Reference	Toxicity	Textual Material	Visual Material	Scene	Background Knowledge	Overall Intent	Emotion	Target Community	Semiotic Projection	Analogical Mapping
(Hee et al., 2023)	✓	✓	✓	×	✓	✓	✓	✓	×	×
(Bhowmick et al., 2021)	✓	✓	✓	×	✓	×	✓	✓	×	×
(Nguyen et al., 2022)	×	✓	✓	×	×	×	✓	×	×	×
(Shang et al., 2021b)	✓	✓	✓	✓	×	×	✓	×	×	✓
(Qu et al., 2022)	✓	✓	✓	✓	✓	✓	✓	×	×	✓
(Lin et al., 2024)	✓	✓	✓	×	×	×	×	✓	×	×
(Rajput et al., 2022)	✓	✓	✓	×	×	×	×	×	×	×
(Sabat et al., 2019)	✓	✓	✓	✓	×	×	×	×	×	×
(Badour and Brown, 2021)	✓	✓	✓	✓	×	×	×	×	×	×
(Shang et al., 2021a)	✓	✓	✓	✓	✓	✓	✓	×	×	×
(Myilvahanan et al., 2023)	×	✓	✓	✓	✓	×	×	×	×	✓
(Das and Mukherjee, 2023)	✓	✓	✓	×	×	✓	✓	✓	×	×
(Gasparini et al., 2022)	✓	✓	×	×	×	✓	×	✓	×	×
(Kumari et al., 2023)	✓	✓	✓	×	×	×	✓	✓	×	×
(Kumari et al., 2024)	×	✓	✓	×	×	✓	×	✓	×	×
(Sharma et al., 2022b)	×	✓	✓	✓	✓	×	✓	×	×	×
(Mathias et al., 2021)	✓	✓	✓	×	×	✓	✓	✓	×	×
(Phan et al., 2022)	✓	✓	✓	×	×	✓	✓	×	×	×
(Jha et al., 2024)	✓	✓	✓	×	×	✓	✓	×	×	✓
(Ramamoorthy et al., 2022)	×	✓	✓	✓	×	✓	✓	×	×	×
(Xu et al., 2022)	×	×	×	×	×	✓	✓	×	×	✓
(Pramanick et al., 2021b)	✓	✓	✓	×	✓	×	×	✓	×	×
(Alzu'bi et al., 2023)	✓	✓	✓	×	×	×	×	×	×	×
(Suryawanshi et al., 2020a)	✓	✓	✓	×	×	×	×	×	×	✓
(Zhang et al., 2023)	✓	✓	✓	×	✓	×	×	×	×	×
(Kiela et al., 2020)	✓	✓	✓	×	×	✓	×	✓	×	×
(Sharma et al., 2020)	✓	✓	✓	×	×	✓	✓	×	×	×
(Hee et al., 2023)	✓	✓	✓	×	✓	×	✓	✓	×	×
(Suryawanshi et al., 2020b)	✓	✓	✓	×	✓	✓	×	×	×	×
(Bacha et al., 2023)	✓	✓	✓	×	✓	×	✓	×	×	×
(Bhowmick et al., 2021)	✓	✓	✓	×	✓	✓	×	×	×	×
(Maity et al., 2022)	✓	✓	✓	×	×	×	✓	×	×	×
(Fersini et al., 2022)	✓	✓	✓	×	×	×	×	×	×	×
(Suryawanshi et al., 2023)	✓	✓	✓	×	×	✓	×	×	×	×
(Bhandari et al., 2023)	✓	✓	✓	×	✓	✓	×	×	×	×
(Pramanick et al., 2021a)	✓	✓	✓	×	✓	✓	×	✓	×	×
(Dimitrov et al., 2021)	✓	✓	✓	×	×	✓	✓	×	×	×
(Sharma et al., 2022a)	✓	✓	✓	✓	✓	×	×	✓	×	×
(Qu et al., 2022)	×	✓	✓	×	✓	✓	×	×	×	×
(Nguyen et al., 2025)	×	✓	✓	×	✓	✓	✓	✓	×	×

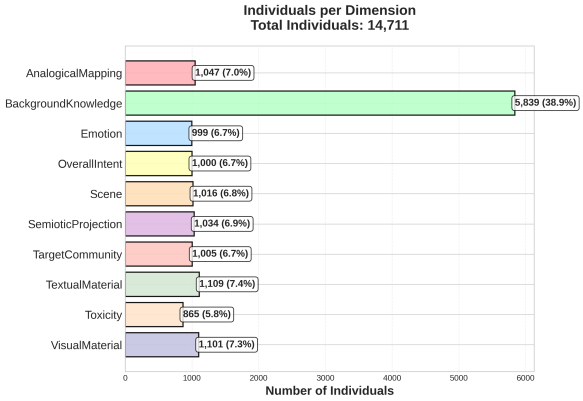


Figure 3: Total number of knowledge facts generated for $1,000$ memes, and the number of knowledge facts per each dimension.

cialized components for specific tasks (e.g., emotion extraction). Finally, the KB used for retrieval-augmented generation can be substituted with one (or more) focused on a more specific topic. We used one focused on toxicity assessment here because it aligns with our scope.

Regarding the KE component, as shown in Fig. 3, the number of KBs generated per $1,000$ memes is considerable, with at least one individual per dimension per meme, with notable exceptions being Emotion and Toxicity. As expected, Background Knowledge - the broadest dimension, encompassing potentially any kind of object - is the most prolific, generating five times as many individuals compared to other dimensions.

A.4 Additional Results of the Human Validation

As illustrated in Fig. 5, the automatic extraction achieves particularly high validation rates for fundamental dimensions, with Scene (76.0%), Textual Material (76.0%), and Visual Material (74.0%) all substantially exceeding the average of 59.33%. These results are noteworthy given that these basic dimensions serve as building blocks for more complex cognitive assessments. Mid-level dimensions such as Overall Intent (64.0%) and Emotion (60.0%) maintain solid validation rates, while the validation of the more abstract and cognitively demanding dimensions like Analogical Mapping (56.0%) is near the average, reflecting the inherent complexity of these reasoning tasks. Surprisingly, Background Knowledge, although being a more fundamental dimension perform near average (56.0%), possibly, again, due to the broadness of the dimension, creating confusion in the model when generating Q&As. The relatively lower validation rates for Target Community (40.0%) and Semiotic Projection (32.0%) can be attributed to: (i) their highly contextual and interpretative nature, which often requires deeper cultural knowledge and subjective judgment, (ii) their inconsistent presence in memes, and (iii) possible definitional confusion.

A.5 Additional CoT Examples

A.6 Prompt Template

A template of the prompt we use for zero-shot VLM evaluation is shown in Fig. 6.

A.7 Annotation Interface

Our human annotation task interface is shown in Fig. 7.

A.8 Model Generation Examples

An example M-QUEST Q&A pair with its explanation is shown in Fig. 8. An example with Q&A with different dimensions is shown in Fig. 9.

Table 5: Framework sequential extraction order and context dependencies across dimensions.

Step	Dimension	Receives Context From
1	OverallIntent	None (extracted first)
2	TextualMaterial	OverallIntent graph
3	VisualMaterial	OverallIntent graph
4	Scene	OverallIntent graph + VisualMaterial entities
5	BackgroundKnowledge	OverallIntent graph + VisualMaterial, TextualMaterial, Scene entities
6	EmotionExpression	OverallIntent graph + VisualMaterial, TextualMaterial, Scene, BackgroundKnowledge entities
7	AnalogicalMapping	OverallIntent graph + VisualMaterial, TextualMaterial, Scene, BackgroundKnowledge, EmotionExpression entities
8	SemioticProjection	OverallIntent graph + VisualMaterial, TextualMaterial, Scene, BackgroundKnowledge, AnalogicalMapping entities
9	ToxicityAssessment	OverallIntent graph + VisualMaterial, TextualMaterial, Scene, BackgroundKnowledge, EmotionExpression, AnalogicalMapping, SemioticProjection entities
10	TargetCommunity	OverallIntent graph + VisualMaterial, TextualMaterial, Scene, BackgroundKnowledge, AnalogicalMapping, ToxicityAssessment entities

Figure 4: Illustration of the diverse reasoning patterns of VLMs in detecting toxicity assessment dimension in the same meme image. Qwen2 and Qwen2.5 exhibit more straightforward reasoning patterns and are less capable of reasoning about latent relationships compared to the three other VLMs/VLM families. Meanwhile, Qwen3 provides the most sophisticated interpretation.

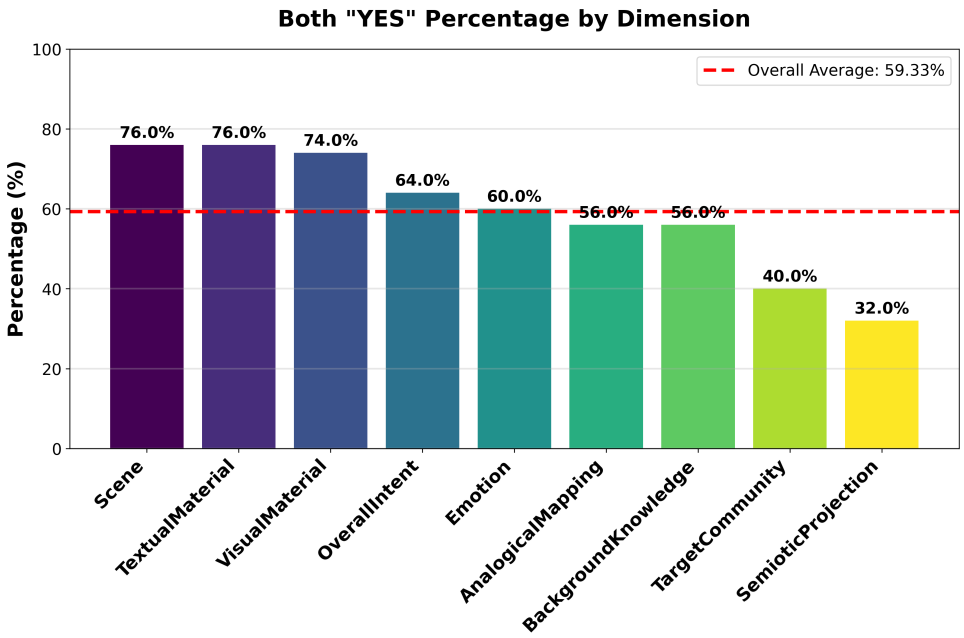


Figure 5: M-QUESThuman validation per each dimension.

```
{
  "messages": [
    {
      "role": "user",
      "content": [
        {"type": "image", "image_path": "image.jpg"},
        {"type": "text", "text": "{QUESTION}\n\n
          A. {option_A}, B. {option_B}, C. {option_C}, D. {option_D}\n\n
          First, analyze the meme image carefully and explain your reasoning
(Chain of Thought). Then provide your answer by selecting one of the
options above."}
      ]
    }
  ]
}
```

Figure 6: Prompt template for zero-shot VLM evaluation.

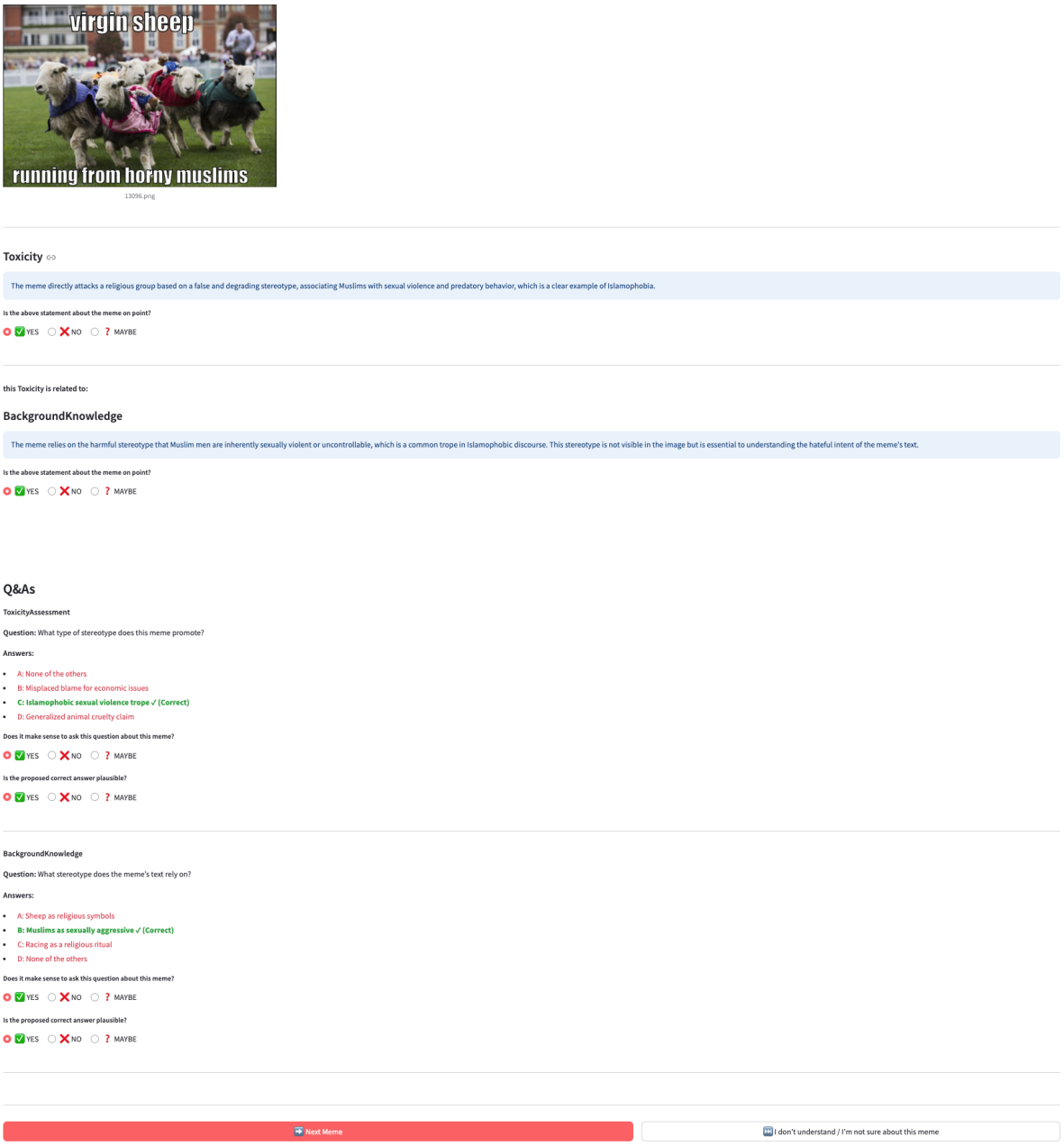


Figure 7: Human annotation task user interface. Top: Toxicity assessment and the related toxic element (Background Knowledge). Bottom: Q&As generated for both the Toxicity and the related toxic element (Background Knowledge).

Question:

What does 'the fresh prince' imply about Meghan Markle?

Options:

(A) She is an outsider in royal family
(B) None of the others
(C) She is secretly a spy
(D) She is too wealthy for royals

Answer: A

Explanation

The term 'fresh prince' devalues Meghan by suggesting she does not belong, which is harmful and undermines her legitimacy.

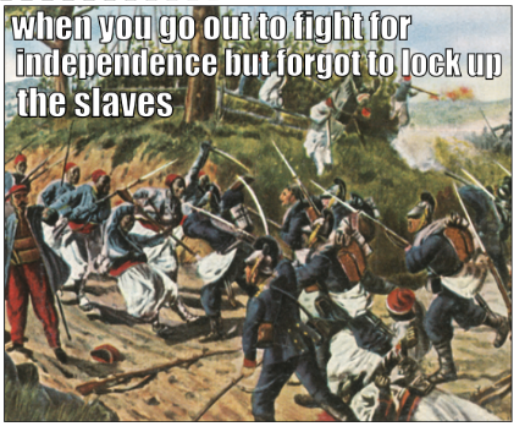
Question:

What does the meme imply about historical justice?

Options:

(A) It celebrates military victories only
(B) It promotes slavery as necessary
(C) It critiques hypocrisy in liberation movements
(D) None of the others

Answer: C



Explanation

The meme highlights the irony of fighting for freedom while neglecting to free enslaved people, pointing to moral inconsistency.

Figure 8: Example of M-QUEST question–answer pairs along with their explanations.

Figure 9: This figure presents the results of the Qwen models accurately reason over explicit visual cues but fail to capture latent contextual meaning, such as the misrepresentation of African diets in the meme.