STEMTOX: From Collaborative Tags to Fine-Grained Toxic Meme Detection via Entropy-Guided Multi-Task Learning

Subhankar Swain* **Naquee Rizwan*** **Vishwa Gangadhar S**
Nayandeep Deb **Animesh Mukherjee**
Indian Institute of Technology, Kharagpur, India
{subhankar.swain25,nrizwan,vishwa2488,nayandeepdeb125}@kgpian.iitkgp.ac.in
animeshm@cse.iitkgp.ac.in
*Equal contribution.

Abstract

Memes, as a widely used mode of online communication, often serve as vehicles for spreading harmful content. However, limitations in data accessibility and the high costs of dataset curation hinder the development of robust meme moderation systems. To address this challenge, in this work, we introduce a first-of-its-kind dataset – **TOXICTAGS** consisting of 6,300 real-world meme-based posts annotated in two stages: (i) binary classification into *toxic* and *normal*, and (ii) fine-grained labelling of toxic memes as *hateful*, *dangerous*, or *offensive*. A key feature of this dataset is that it includes *collaborative tags* associated with the original posts, enhancing the context of each meme. In addition, we propose a novel entropy-guided multi-tasking framework – **STEMTOX** – that leverages these collaborative tags alongside visual and textual inputs within a robust classification framework. Experimental results show that incorporating these tags substantially enhances the performance of state-of-the-art VLMs in toxicity detection tasks. Our contributions offer a novel and scalable foundation for improved content moderation in multimodal online environments. We have made our code[1] and dataset[2] publicly available for research purposes. ***Warning: Contains potentially toxic contents.***

1 Introduction

While communication on online platforms span a variety of modalities[3], memes have emerged as a popular and influential form of expression – initially intended for lighthearted humor. However, they are increasingly being misused as vehicles for spreading harmful content, including hate speech, misinformation, and toxic ideologies.

Model limitations: Identifying such harmful content is particularly challenging due to the subtle and context-dependent nature of memes. Their meaning is often embedded in cultural references, online trends, sarcasm, or coded language, making them difficult to interpret not only for automated systems but even for human moderators. In response to these challenges, vision language models (VLMs) have recently gained traction as powerful tools for content moderation. These models are capable of jointly analysing visual and textual elements to grasp the nuanced context of memes and can provide detailed justifications for their classifications (Qu et al., 2025). Despite these advances, recent studies have highlighted the limitations of VLMs in accurately detecting hateful memes (Rizwan et al., 2025a), underscoring the urgent need to bridge these performance gaps. Similarly, a recent blog post[4] by Meta highlights the challenges and complexities inherent in their current content moderation frameworks.

Dataset scarcity: Datasets serve as the foundational fuel for generative AI models. However, researchers increasingly face significant obstacles in curating such datasets, primarily due to the high costs of manual annotation and the limitations imposed on large-scale crawling of social media platforms. These challenges have resulted in datasets that are either manually constructed (Kiela et al., 2020; Lin et al., 2026) – often failing to fully capture the richness of human creativity – or narrowly scoped to specific targets or events (Pramanick et al., 2021a,b; Fersini et al., 2022). Moreover, this scarcity of comprehensive datasets has hindered efforts to build robust toxicity detection systems for multimodal social media content.

[1]Code: https://github.com/hate-alert/STEMTOX

[2]Dataset: https://huggingface.co/datasets/swainsubhankar/ToxicTags

[3]https://www.sprinklr.com/blog/types-of-social-media/

[4]https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

Transactions of the Association for Computational Linguistics, vol. 14, pp. 2281–2300, 2026. https://doi.org/10.1162/TACL.a.801
Action Editor: Sebastian Padó. Submission batch: 2/2026; Revision batch: 6/2026; Published 8/2026.

This is in contrast to textual hate speech research, where more extensive taxonomical explorations exist (Sachdeva et al., 2022).

Our contributions: **(A)** We introduce a diverse real-world memes dataset (**TOXICTAGS**) that incorporates collaborative tags collected from associated post metadata– critical yet often overlooked feature in social media content that is extensively used for describing, categorising, or commenting on digital content[5] . For instance, tags such as *#hitler, #jews, #muslim, #alabama, #horse_meat,* and similar others provide contextual grounding for the corresponding memes. Unlike prior datasets, it consists of *only* real-world memes with no restrictions based on specific targets or events (see Section 3). The final dataset comprises **6,300** annotated memes, capturing a wide spectrum of online discourse. **(B)** We design a rigorous two-stage human annotation pipeline. In the first stage, memes are annotated as either *toxic* or *normal*. In the second stage, *toxic* memes are further categorised into one of the three fine-grained classes: *hateful*, *dangerous*, or *offensive*. These categories have been distilled through an iterative annotation process, revealing that a four-class taxonomy is appropriately expressive for moderating a wide range of social media content. This taxonomy can help reduce misclassifications in existing toxicity detection pipelines. **(C)** We introduce **STEMTOX**, a novel framework that significantly advances state-of-the-art by generating high-quality collaborative tags and improving meme classification performance. The overall architecture of the proposed framework is presented in Figure 1 and Section 6.

Key results

☞**STEMTOX** achieves (refer Table 4) the best overall performance, outperforming several approaches in the classification task with a Stage I macro-F1 score of **72.55%** and a Stage II macro-F1 score of **64.66%**; improvements of **2.07%** and **3.12%**, respectively over the most competing baseline. In addition, it simultaneously generates high-quality collaborative tags with scores of **0.4872** (CHRF), **0.265** (METEOR), and **0.626** (SEMANTIC-SIMILARITY), again outperforming the competing approaches. Our experiments demonstrate that incorporating collaborative tags into the multitasking system leads to significant performance gains in meme classification tasks.

☞The classification performance of our method on two other popular benchmark hate meme detection datasets compared to four prior approaches demonstrates the generalizability of our method. As presented in Table 5, **STEMTOX** outperforms the strongest baseline by **3.30%** and **3.26%** in terms of accuracy and macro-F1 on **FHM**, and by **6.47%** and **6.25%** on **MAMI**, which shows its effectiveness in various meme classification settings.

☞ Finally, for the tag generation task, we compare our method with three prior approaches. As illustrated in Table 6, **STEMTOX** achieves improvements of **0.277** in CHRF, **0.178** in METEOR, and **0.294** in SEMANTIC-SIMILARITY.

2 Related works

Hate meme detection: Rapid surge of hate around the globe (Arcila Calderón et al., 2024) in the recent past years, with memes acting as a major source of fuel, has led to the curation of multiple hateful memes datasets (Kiela et al., 2020; Lin et al., 2026). Findings from prior work on MLLMs (Wang et al., 2025) underscore the challenges in robust multimodal safety understanding. Similarly, there has been extensive research in the field to build robust content moderation frameworks (Rizwan et al., 2025a; Mandal et al., 2024; Yerukola et al., 2025; Das and Mukherjee, 2023b; Prakash et al., 2023; Cao et al., 2024a, 2022), with some works on low-resource languages (Das and Mukherjee, 2023a; Kumari et al., 2024; Xue et al., 2026). Some works are extended to audio and video modalities (Boishakhi et al., 2021; Rizwan et al., 2025b; Koushik et al., 2026) as well. Despite such rapid developments, current datasets are either limited to manually curated memes or focus only on a subset of events (Pramanick et al., 2021a,b; Chen et al., 2023).

Tags: Prior research has extensively explored automated tag generation for images (Huang et al., 2024; Zhang et al., 2024; Dai et al., 2023), leading to the development of several dedicated models and datasets. Such synthetic tags have also been used for hate meme detection recently (Koushik et al., 2026). In all the above case, tags refer to general categorizing information used to support resource organization, while collaborative tagging refers to the practice of describing, categorizing, or commenting on digital content by users[6]. How-

[5] https://www.imrpress.com/journal/ko/45/6/10.5771/0943-7444-2018-6-500

[6] https://www.imrpress.com/journal/ko/45/6/10.5771/0943-7444-2018-6-500, https://www.isko.org/cyclo/tagging

dataset	event / target independent?	real world memes?	post's contextual information?	two stage annotation?	size	labels
FHM (Kiela et al., 2020)	✓	✗	✗	✗	∼10k	hateful, not-hateful
MAMI (Fersini et al., 2022)	✗	✓	✗	✗	∼10k	misogynistic, not-misogynistic
HARM-C (Pramanick et al., 2021a)	✗	✓	✗	✗	∼3,544	very/partially harmful, harmless
HARM-P (Pramanick et al., 2021b)	✗	✓	✗	✗	∼3,552	very/partially harmful, harmless
UA–RU Conflict (Thapa et al., 2022)	✗	✓	✗	✗	∼5,680	hateful, not-hateful
CrisisHateMM (Bhandari et al., 2023)	✗	✓	✗	✗	∼4,723	hateful, not-hateful
RUHate-MM (Thapa et al., 2024)	✗	✓	✗	✗	∼20,675	hateful, not-hateful
TOXICTAGS	✓	✓	✓	✓	∼6,300	stage I: toxic, normal stage II: hateful, dangerous, offensive, normal

Table 1: Comparison of **TOXICTAGS** with existing datasets for toxic meme detection. Post-contextual information includes metadata such as titles or tags associated with the meme.

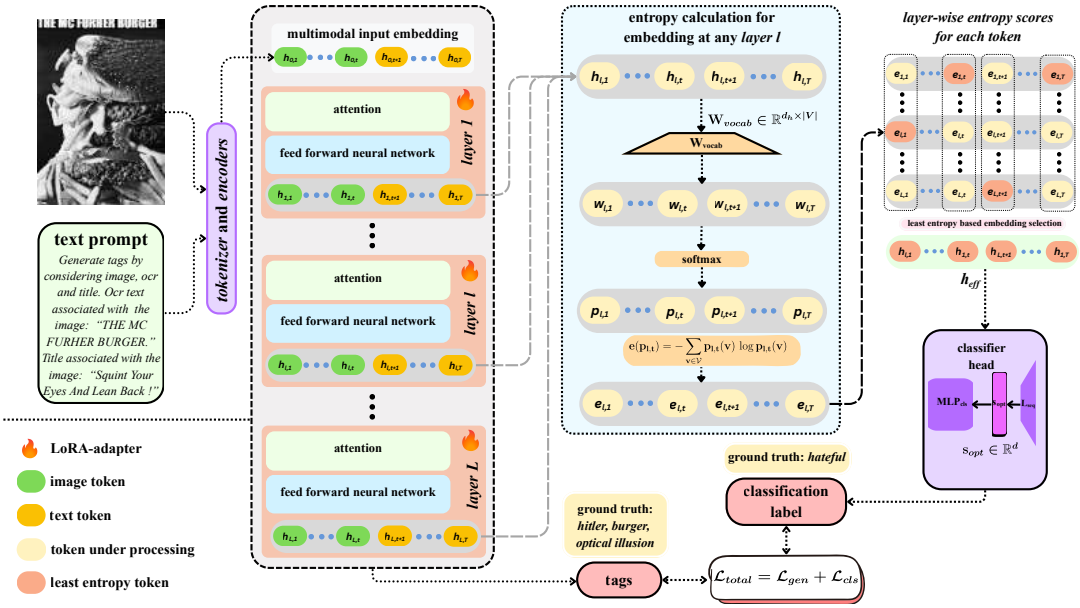


Figure 1: Proposed **STEMTOX** framework. Refer to Section 6 for the architectural overview.

ever, to the best of our knowledge, no existing work addresses the specific challenge of tag generation for memes – visual artefacts that often combine text and imagery in culturally nuanced, context-dependent ways.

Present work: We present a real-world memes dataset (**TOXICTAGS**) with fine-grained toxicity labels beyond binary classification (Table 1). In addition, each meme is accompanied by collaborative tags obtained from associated post metadata. We also introduce an entropy-based multi-task learning framework **STEMTOX** for open-set image tagging and robust classification.

3 The TOXICTAGS dataset

Collection: We source real-world social media memes from [https://imgflip.com], specifically utilising its *streams* section[7] to crawl meme-centric posts. The platform was chosen for two primary reasons: **(i)** its strong emphasis on social media-style interactions where memes serve as a central mode of communication, and **(ii)** its wide variety of user-generated content, allowing us to collect memes without any target-specific or event-specific filtering, thereby closely mirroring real-world social media environments.

For our dataset, we select three high-volume and thematically rich streams—

[7] https://imgflip.com/streams

DARK_HUMOUR[8], MEMES_OVERLOAD[9], and POLITICS[10]—based on their popularity and relevance to diverse meme content. We selected these streams through manual inspection of their descriptions and content samples (see Appendix C for details). Further, to ensure that the memes included are socially engaging and contextually rich, we manually review a subset of posts and observe that memes with fewer than two comments are often unengaging or irrelevant to a broader audience. Hence, we retain only those memes that received at least two comments. In total, we manually collect 37,072 memes across the selected streams over a period of approximately one month. The memes were then downloaded using a browser-based image downloader extension[11]. As a result, our dataset offers a large and diverse collection of real-world memes, free from artificial filtering or event-driven bias.

Metadata: For each post, along with the meme, we collect its **(i)** title, **(ii)** number of comments and **(iii)** the tags list. For the collected memes, we extract the embedded text using GOOGLE OCR[12]. To ensure robustness of OCR-extracted text, we perform manual verification of 100 randomly chosen samples and find 98 to be correctly identified, thus depicting the robustness of the tool.

Pre-processing: Before providing the meme along with its metadata for annotation, we perform the following pre-processing steps.

(i) *Deduplication*: Out of the initially collected 37,072 posts, we first apply exact string matching on the `title` and `tags` fields to identify duplicate entries. Subsequently, we perform visual deduplication on the matched posts using the IMAGEDEDUP[13] library, employing a Hamming distance threshold of zero to ensure strict removal of visually identical images. This two-step deduplication process – textual followed by visual – ensures that only unique memes, along with their associated metadata, are retained. After this filtering, we obtain a final dataset comprising 6,300 unique and high-quality memes.

[8] https://imgflip.com/m/Dark_humour

[9] https://imgflip.com/m/MEMES_OVERLOAD

[10] https://imgflip.com/m/politics

[11] https://chromewebstore.google.com/detail/download-all-images/ifipmflagepipjokmbdecpmjbibjnakm?hl=en

[12] https://docs.cloud.google.com/vision/docs/ocr

[13] imagededup documentation

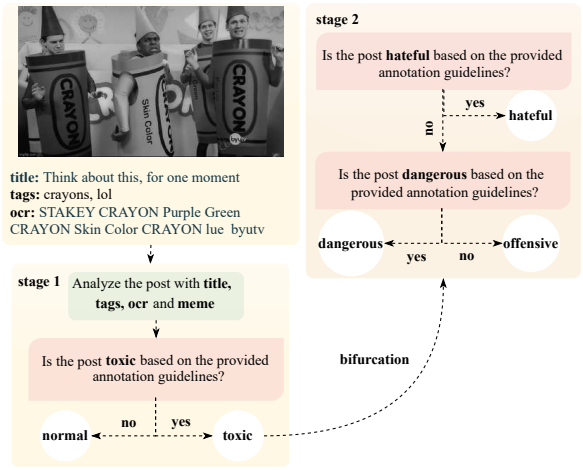


Figure 2: A step-by-step flowchart used by annotators to annotate any post.

(ii) *Removal of unwanted tags*: To reduce noise and enhance the quality of our tags, commonly occurring irrelevant tags like – *'darkhumour', 'memes', 'you have been eternally cursed for reading the tags'* – are manually removed from each post by expert researchers (refer to Section 4). Subsequently, we obtain a rich set of 7,209 unique and collaborative tags from an initial list of 9,664 unique tags.

Agreement: We strictly adhere to the privacy and copyright regulations of the platform[14] and collect our data only from the publicly available posts. To maintain the privacy of users, we did not store any information that could potentially compromise their anonymity.

4 Annotation

Two-stage annotation: We follow a two stage annotation process – **(i)** classification of each meme into *toxic* or *normal*, and **(ii)** further bifurcation of toxic memes into *offensive*, *dangerous* or *hateful*. The two-stage process is specifically adopted to mitigate annotation errors as suggested in (Rizwan et al., 2025a). We agree that hateful memes can be both offensive and violent, but not vice versa. Hateful memes target a particular community and impose threat and offense upon them. This distinguishes hateful memes from violent and offensive memes. Similary offensive content is limited to abusive or derogatory language, while dangerous content involves expressions that may increase the risk of violence against a someone (for definitions refer to Appendix D.1).

[14] https://imgflip.com/terms

Prior works also suggest (Mathew et al., 2021; Davidson et al., 2017) that hate classification is a subjective task, and highlight the importance of distinguishing between closely related but conceptually different forms. Following these prior works, we adopt a mutually exclusive labeling scheme, where each instance is assigned a single primary category. This design is also consistent with real-world content moderation systems, where most streaming and social media platforms (e.g., YouTube[15], Facebook[16]) typically allow users to select only one primary violation category. Figure 5 in the Appendix presents snapshots of the policy/reporting interfaces of Facebook, Youtube, and LinkedIn, where users are generally guided to select a single primary violation category.

In subjective tasks such as hate speech or toxicity detection, annotator agreement alone does not necessarily guarantee validity due to variability in lifestyle, culture, and demographic backgrounds. Prior work (Röttger et al., 2022) distinguishes between the prescriptive paradigm, which enforces a consistent labeling standard, and the descriptive paradigm, which captures diverse annotator beliefs. In this work, we adopt a **prescriptive annotation paradigm** to ensure consistency and reduce ambiguity in labeling. During the annotation process, annotators were instructed to strictly adhere to the label definitions.

Figure 2 presents a brief pipeline of the employed annotation process over two stages. The instructions that we designed are given below, and some samples from those provided to annotators are present in Table 10 in the Appendix. Further, we also conducted a pilot study at each stage before progressing to the final annotations. For both stages, we perform **three annotations** per sample by three different annotators to minimise annotation bias. We use the standard definition of labels that are effectively used for practical applications. In the upcoming subsections, we discuss each stage separately and present further detailed annotation instructions with corresponding definitions of the labels in Appendix D.

Annotator details: To ensure a high-quality annotation process, we shortlisted 25 annotators based on the following minimum criteria: **(i)** candidates having professional experience and prior work aligned with our research, and **(ii)** familiarity with diverse social media content. The pool included 14 male and 11 female participants. To ensure high-quality annotations, we selected 19 Master's students and 6 PhD students aged between 22 and 40 years. In addition to the above, the group included diverse religious backgrounds: 14 identified as Hindus, 8 as Muslims, and 3 as Christians. We additionally filtered annotators by asking them to pass a qualification task curated by the supervising researchers to assess label comprehension and consistency. The entire annotation process was conducted under the close supervision of two PhD researchers. We detail the safety guidelines in Appendix D; further, as an extra safety measure, we rolled out a maximum of only 50 samples per day to each annotator.

Annotation codebook: We also provide detailed annotation instructions, accompanied by 25 samples that have been manually selected and annotated by the two supervising researchers. As outlined in Appendix D, our annotation codebook has been constructed based on standard guidelines from which we have derived our definitions. For the toxic label, we strictly follow the definition used by the Perspective API, given its widespread adoption in production settings. For the hateful label, we adhere to the Facebook hate meme definition (Kiela et al., 2020), which has become the de facto standard. For the dangerous label, we follow the annotation guidelines available at the following source[17].

4.1 Stage I: Binary labelling of the dataset

Annotation: The full set of 6,300 samples has been evenly distributed among the 12 selected annotators (out of the 25 recruited). As stated earlier, we employ three annotators per sample, which enhances annotation diversity. We achieve a final inter-annotator agreement score of 0.753, as measured by Fleiss' κ. Notably, for a subjective task (Mathew et al., 2021) such as ours, this level of agreement is considered relatively high, thereby validating the quality of our annotations. Each annotator has been paid USD 40 for this task, which is much above the minimum wage in the annotators' country.

Label assignment: The final label for each sample is determined based on the majority vote of the

[15] https://support.google.com/youtube/answer/9288567

[16] https://www.facebook.com/privacy/center

[17] https://www.dangerousspeech.org/dangerous-speech

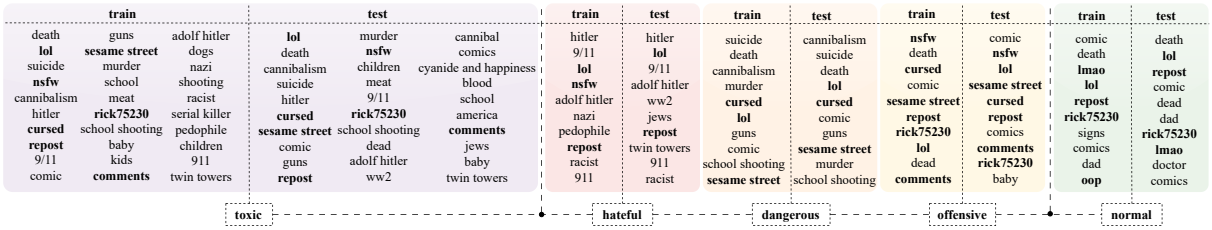


Figure 3: Frequently occurring tags in the *toxic*, *hateful*, *offensive*, and *normal* classes.

three annotations, i.e., the label with at least two agreements is considered to be the final label. In this stage, we finally obtained 1,446 normal and 4,854 toxic samples (refer to Table 2 for complete statistics).

4.2 Stage II: Fine-grained labelling

Annotation: As a pilot step, we randomly sampled 250 toxic memes from the previous stage and had the expert annotators categorise them into three distinct labels. These labelled samples were then evenly distributed among the 12 annotators, each of whom annotated all the assigned samples. Upon verification, we found that the annotators were performing satisfactorily and thus proceeded with the final annotation with this group. Subsequently, the set of 4,854 toxic samples was categorised into three classes: *hateful*, *dangerous*, and *offensive*. As before, each sample received three independent annotations. The annotators achieved a Fleiss' κ agreement score of 0.793 – a slight improvement over Stage I – highlighting the value of multi-stage and multi-pilot studies in enhancing annotation quality.

Label assignment: The final label is determined based on the majority vote of the annotators for each sample. In total, 26 samples were identified as undecidable cases, where majority voting failed. These undecidable instances mainly include posts containing mixed linguistic cues, subtle contextual dependencies or the coexistence of multiple possible categories. Given the relatively small number of such instances, final annotations for these cases were determined by expert researchers following the same annotation guidelines to maintain consistency. The overall dataset statistics are noted in Table 2. To maintain the diversity of tags in both the splits, we ensure that each tag appears at least 15% of the time in the test split, resulting in a total of 1,000 test samples and 5,300 training samples. Word clouds for label-wise tags are presented in Table 11 of Appendix E.

stage	label	train	test	total
I & II	normal	1297	149	1446
I	toxic	4003	851	4854
II	hateful	1470	282	1752
	dangerous	1847	472	2319
	offensive	686	97	783
total		5300	1000	6300

Table 2: Dataset statistics showing the distribution of samples across classes.

tag pairs	theme
('hitler', 'nazi'), ('hitler', 'jews'), ('hitler', 'ww2'), ('hitler', 'holocaust'), ('jews', 'nazi'), ('concentration camp', 'nazi'), ('hitler', 'stalin')	Historical events / World War II
('9/11', 'twin tower'), ('911', 'twin tower'), ('9/11', 'muslim'), ('cow', 'muslim'), ('911 9/11 twin tower impact', 'muslim')	9/11 and Islamophobic references
('jesus', 'satan'), ('satan', 'the bible'), ('jesus', 'the bible')	Religious contrast / satire
('atomic bomb', 'hiroshima'), ('hiroshima', 'ww2'), ('hiroshima', 'nuke')	Nuclear warfare / World War II
('alabama', 'incest'), ('incest', 'sweet home alabama'), ('alabama', 'dad')	Southern US stereotypes, taboo, humour
('cannibal', 'cannibalism'), ('fresh', 'meat'), ('cannibalism', 'meat'), ('human', 'meat')	Creepy / disturbing themes
('mass shooting', 'school shooting'), ('gun', 'school shooting'), ('gun', 'school'), ('gun', 'usa'), ('mass shooting', 'usa'), ('mass shooting', 'america')	US gun violence
('baby', 'dead'), ('baby', 'cannibalism')	Child-related violence / abortion
('africa', 'water'), ('africa', 'hungry'), ('africa', 'starve')	Humanitarian crises in Africa
('black people', 'racist'), ('black', 'black life matter'), ('black', 'white'), ('angry black guy', 'lame')	Racism
('elmo', 'sesame street'), ('big bird', 'sesame street'), ('ernie', 'sesame street'), ('mayor', 'serial killer'), ('free candy van', 'sonic the hedgehog')	Cartoon / anime characters in dark contexts
('world war 3', 'ww3'), ('ukraine', 'ww3'), ('gaza', 'israel'), ('monster', 'ww3')	Global conflict / war escalation

Table 3: Top co-occurring tag pairs and associated semantic themes.

5 Analysis of the dataset

One of the most unique features of our dataset is the tags associated with each meme. This allows us to perform certain interesting analyses that we report below.

Frequent tags: Some of the most frequent tags we find in the train and the test splits are illustrated in Figure 3. We report the top 30 tags from the toxicity class as well as the top 10 tags each from the hateful, dangerous, offensive and normal classes, respectively. The key observations are as follows.

(i) The distribution of the top tags across the train and the test splits is very similar, indicating that our strategy of splitting the data is effective.

(ii) The most prevalent tags in the *hateful* class are *'9/11', 'nazi', 'adolf hitler', 'paedophile'* and *'racist'*, indicating their popularity in spreading hateful sentiments. We also observe that for the *dangerous* class, some tags that appear benign at the surface level and are primarily associated with children's media show up. These include *'ernie and bert', 'sesame street'*, and *'sonic the hedgehog'*, which are used to render a comic angle to the dangerous posts (see Table 12 in the Appendix for more examples).

(iii) We observe that users are generally reluctant to reshare content involving serious violence, suicide, or murder, as indicated by the relatively low frequency of the *'repost'* tag in the *dangerous* category compared to its prevalence in other categories.

(iv) The widespread use of the *'lol'* tag suggests an attempt to downplay harmful content through humour.

(v) Frequent use of tags such as *'cannibalism', 'murder', 'suicide'*, and *'abortion'* and violent phrases like *'school shooting'* underscores how dark humour is often employed as a tool to normalise or propagate digital violence, thereby contributing to a heightened sense of danger.

(vi) One notable observation concerns the use of the tag *'alabama'*, which frequently appears in hateful memes that mock familial relationships, highlighting how geographic stereotypes are weaponised for humour and hate.

Related tags: To understand the semantic themes associated with frequently occurring tag pairs, we conduct co-occurrence analysis. Prior to the analysis, all tags are lemmatised to ensure consistency and improve result accuracy. Our primary objective is to identify how often specific tags co-occur and what thematic contexts they represent within the meme content. Some of the top co-occurring pairs and their themes are listed in Table 3. The co-occurring tag pairs in the table indicate that toxic meme content is often structured around semantically linked concepts. For example, themes such as *Historical events / World War II* frequently associate tags like *'hitler'* and *'nazi'*. Similarly, in the context of *9/11 and Islamophobic references*, tags such as *'twin tower', '9/11'*, and *'muslim'* often co-occur. Further, tags like *'africa'* and *'water'* are grouped under themes related to *Humanitarian crises in Africa*. These patterns suggest that toxicity is reinforced through well-established contextual associations. The theme of *Cartoon / anime characters in dark contexts* indicates that such characters often co-occur with disturbing or violent concepts, suggesting that they are used to amplify humour or shock value. Overall, this analysis highlights that meme semantics inherently arises from the interaction of the multiple collaborative tags.

6 The STEMTOX framework

6.1 Entropy-guided layer selection objective

Past approaches (Cao et al., 2022, 2023, 2024b; Hee et al., 2024; Lu et al., 2025) in toxic meme detection have focused primarily on single-task setups and rely either on zero-/few-shot prompting or perform low-resource fine-tuning using strategies like PEFT/LoRA. Recently, multimodal research has evolved, and some works like (He et al., 2025; Luo et al., 2024) propose a deeper understanding of the model's internal representations and semantic alignment. Recent empirical evidence presented in (Chen et al., 2025) suggests that deep representational alignment between vision and language modalities peaks at intermediate depths before undergoing representational drift in terminal layers. Taking cues from these works, we propose a novel token confidence-based approach for simultaneous tag generation and toxic meme classification. Below, we detail our approach.

Figure 1 illustrates the complete workflow of STEMTOX. We consider a multimodal model $\mathcal{M}$ that takes a text prompt $\mathcal{T}$ and an image $\mathcal{I}$ as input and consists of L stacked layers. Given an input sequence of fixed length, the model produces hidden representations across layers. Let $l \in \{1, \ldots, L\}$ (including the token embedding

layer) denote the layer index, $t \in \{1, \ldots, T\}$ denote the token position in the input sequence, and $v \in V$ denote a token from the model vocabulary V. The hidden state $\mathbf{h}_{l,t} \in \mathbb{R}^{d_h}$ corresponds to token t at layer l, as depicted in the figure, where d_h represents the hidden dimension of the model. To process the model's internal state, we extract token embeddings from all layers and project them into the model's vocabulary space using the pretrained language model head $\mathbf{W}_{\text{vocab}} \in \mathbb{R}^{d_h \times |V|}$. This projection produces $\mathbf{w}_{l,t}$, which indicates the representation of the t-th token at layer l in the vocabulary V:

$$\mathbf{w}_{l,t} = \mathbf{h}_{l,t}\, \mathbf{W}_{\text{vocab}}. \tag{1}$$

A softmax function is applied to $\mathbf{w}_{l,t}$ to obtain the confidence distribution $p_{l,t}$ over the vocabulary for the t-th token at layer l:

$$p_{l,t} = \text{softmax}(\mathbf{w}_{l,t})\,, \tag{2}$$

To measure the confidence of the representation at layer l for token t, we compute the Shannon entropy (Vajapeyam, 2014) $e_{l,t}$, which represents the entropy of token t at layer l, as follows:

$$e_{l,t} = -\sum_{v \in V} p_{l,t}(v)\, \log p_{l,t}(v). \tag{3}$$

This entropy measure enables the dynamic selection of the least random token embedding across layers by selecting the representation corresponding to the minimum entropy, enabling the classification head to extract features from the most certain internal representations $\mathbf{h}_{\text{eff}}$:

$$\mathbf{h}_{\text{eff}} = \text{Aggregate}\left(\left\{ \mathbf{h}_{l^*,t} \;\middle|\; l^* = \arg\min_{l \in \mathcal{L}} e_{l,t} \right\}_{t=1}^{T} \right). \tag{4}$$

Finally, the confident sequence representation $\mathbf{h}_{\text{eff}}$ is passed through the classification head to predict the class logits. The classification head consists of a sequence compression layer that compresses the sequence and projects it to a fixed-dimensional representation, which is then passed through an Multi-Layer Perceptron(MLP) for final classification:

$$\text{classification label} = \text{MLP}(\text{Compress}(\mathbf{h}_{\text{eff}})) \tag{5}$$

We employ a weighted cross-entropy loss for the classification task, which is depicted as $\mathcal{L}_{\text{cls}}$ in the figure. To address class imbalance, class weights are computed following the weighting scheme described in (King and Zeng, 2001)[18].

6.2 Multitask learning objective

A meme might not always be accompanied by tags, unlike in the case of our dataset. It is therefore important to have an automatic method to assign tags to an arbitrary input meme, which could enhance toxicity detection. As a combined objective, we therefore propose a novel multitasking framework where we jointly train toxic tag generation and discriminative label classification. The main motivation of this framework is to leverage semantic supervision to guide discriminative classification, where the embeddings are shaped into a representation space where the supervision provided by the generative task acts as a regulariser for the shared representation space. Our experimental observations (refer to Section 7) reveal that supervising the model to generate explicit toxic tags stimulates the intermediate hidden representations of the model and boosts discriminative performance, thereby signalling that the model learns more detailed and meaningful representations in joint training. The overall loss is calculated as a combination of the generation and classification losses:

$$\mathcal{L}_{total} = \mathcal{L}_{gen} + \mathcal{L}_{cls} \tag{6}$$

where, $\mathcal{L}_{gen}$ represents the generative loss and $\mathcal{L}_{cls}$ represents the classification loss. This alignment makes the layer selection objective described in Section 6.1 more robust as a toxicity classifier. Specifically, $\mathcal{L}_{gen}$ is the default token-level cross-entropy loss over the generated tag sequence, whereas $\mathcal{L}_{cls}$ is defined as a weighted cross-entropy loss for the classification task. The details on the employed models and experimental setup are elaborated in Appendix F and G, respectively. In contrast, the uni-tasking framework (see Table 4), considers only the classification loss $\mathcal{L}_{\text{cls}}$, since it is solely dedicated to toxicity detection. In the next section, we detail our empirical observations.

7 Results

In this section we assess the performance of **STEMTOX** toxic meme detection on the test split of the **TOXICTAGS** dataset for both stages – I

[18] scikit-learn documentation

approach	model	shot	mF1↑		tag similarity↑		
			STAGE I	STAGE II	CHRF	MET	SS
multimodal prompting							
zero-shot	GPT	0	54.57	41.62	–	–	–
	LLAVA	0	45.84	10.19	–	–	–
few-shot *(random)*	GPT	2	61.09	43.44	–	–	–
		4	62.46	44.07	–	–	–
	LLAVA	2	49.28	21.32	–	–	–
		4	51.50	24.26	–	–	–
few-shot *(tag similarity)*	GPT	2	63.96	50.48	–	–	–
		4	64.86	52.88	–	–	–
	LLAVA	2	59.99	40.62	–	–	–
		4	59.71	41.48	–	–	–
few-shot *(image similarity)*	GPT	2	61.86	48.59	–	–	–
		4	64.89	50.98	–	–	–
	LLAVA	2	54.62	34.57	–	–	–
		4	55.05	35.66	–	–	–
unimodal prompting							
zero-shot	GPT	0	46.72	30.22	–	–	–
	LLAMA	0	48.07	27.17	–	–	–
	MISTRAL	0	55.48	31.60	–	–	–
baselines							
PH (Cao et al., 2022)	–	–	61.83	–	–	–	–
PC (Cao et al., 2023)	–	–	63.26	–	–	–	–
MH (Cao et al., 2024b)	–	–	49.84	–	–	–	–
BM (Hee et al., 2024)	–	–	59.70	–	–	–	–
ablations: last hidden layer							
uni-task	LLAVA	–	61.58	44.69	–	–	–
multi-task	LLAVA	–	65.03	54.56	0.2585	0.1050	0.4540
uni-task	PALIGEMMA	–	61.52	45.92	–	–	–
multi-task	PALIGEMMA	–	70.48	61.54	0.4782	0.2600	0.6220
proposed: STEMTOX							
uni-task	LLAVA	–	62.19	51.90	–	–	–
multi-task	LLAVA	–	67.24	57.99	0.2648	0.1070	0.4470
uni-task	PALIGEMMA	–	60.85	42.78	–	–	–
multi-task	PALIGEMMA	–	**72.55**	**64.66**	**0.4872**	**0.2650**	**0.6260**

Table 4: Comparative results for toxicity detection and tag similarity. Best results are **highlighted.** **Abbrev.:** PH: PROMPTHATE; PC: PROCAP; MH: MODHATE; BM: BRIDGING MODALITIES; mF1: Macro-F1; MET: METEOR; SS: SEMANTIC-SIMILARITY; – denotes unsupported tasks.

and II. We also present the results of tag generation performance achieved by **STEMTOX**. For our experiments, we use two open-source VLMs – PALIGEMMA and LLAVA (refer Appendix F and G for further details). Table 4 shows the experimental results.

Toxic meme detection: **STEMTOX** with the multi-task learning objective, it consistently outperforms uni-task training across both models and stages, highlighting the benefit of joint learning for toxicity detection and tag generation, as summarised in Table 4. PALIGEMMA attains the high-

est macro F1 scores of 72.55 and 64.66 for stages I and II, respectively. This represents a notable improvement over the uni-task setting, which has macro F1 scores of 60.85 in stage I and 42.78 in stage II. A similar pattern is also observed for LLAVA, where **STEMTOX** in the multi-task framework achieves macro F1 scores of 67.24 and 57.99 for the two stages, respectively, outperforming the uni-task learning approach. Likewise, **STEMTOX** also outperforms the setup that uses the last hidden layer to compute $\mathcal{L}_{cls}$ instead of the proposed entropy-based scheme.

Zero-shot and few-shot variants: Notably, **STEMTOX** outperforms several zero-shot and few-shot variants illustrated in Table 4. Below, we provide a brief overview of these methods (for prompts, refer to Appendix B). To assess the pretrained capabilities of the popular VLM models, we perform zero-shot evaluation using LLAVA and the powerful GPT by giving input as an image, text and title associated with the image. Although PALIGEMMA is used as the backbone of our framework, it has been red-teamed and cannot be reliably used for hate classification without task-specific fine-tuning. We also investigate whether the tags associated with a meme are themselves sufficent for prediction (i.e., the meme image is not part to the input). To test this, we consider pretrained LLM models – MISTRAL, LLAMA and GPT and perform the prediction by providing a prompt having only the tags associated with the meme as input.

We further investigate, under consistent model settings, few-shot variants with diverse sampling strategies to ensure a fair comparison. Few-shot sampling strategies are as follows. (i) *Few shot with random sampling*: In this setting, samples are chosen randomly with different seed values such as 19, 42, and 97. (ii) *Few shot with image similarity*: Instead of selecting random shots, we select those memes as few-shot examples that are most similar to the query meme in terms of the similarity between their CLIP-based image embeddings. **STEMTOX** outperforms all these zero-/few-shot variants as observed in Table 4. (iii) *Few shot with tag similarity*: The few-shot examples are selected based on the ground-truth tag similarity of the memes. To obtain these few-shot samples, we first compute the CLIP embeddings for each tag in the train set and also those that are associated with the query sample. Next, we find the maximum cosine similarity of each query tag embedding with the tag embeddings of an example meme from the train set. For instance, given a test sample with tags $\{t_1, t_2\}$ and an example meme with tags $\{e_1, e_2, e_3\}$, we compute the cosine similarity between the CLIP embedding of tag t_1 and those of e_1, e_2, and e_3, and record the maximum similarity value. We repeat the same process for t_2. The final tag similarity between the test and the example sample is then defined as the mean of the maximum similarity scores.

Competing baselines and generalizability across datasets: To further ground **STEMTOX**, we compare it with the existing SOTA methods. We compare these methods using **TOXICTAGS** as well as two well-studied benchmark datasets (**FHM** and **MAMI**). The methods that we choose are PROMPTHATE (Cao et al., 2022), PROCAP (Cao et al., 2023), MODHATE (Cao et al., 2024b) and BRIDGING MODALITIES (Hee et al., 2024).

Table 4 shows that **STEMTOX** demonstrates substantial improvement over all these existing approaches for the **TOXICTAGS** dataset. In case of **FHM** and **MAMI** datasets as benchmarks, we do not have tags that is necessary to train **STEMTOX**. For the training points of both these datasets we therefore infer the tags using a PALIGEMMA model that is fine-tuned on the **TOXICTAGS** dataset to predict tags. Next, we use the gold labels and these inferred silver tags to train **STEMTOX** (PALIGEMMA backbone, as it exhibits the best performance in Table 4) for each of the two datasets. Table 5 compares the results of the baselines with **STEMTOX**. Our approach shows substantial improvement over the other baselines for the **FHM** and **MAMI** datasets, achieving a macro F1 score of 78.11 and 79.67 and an accuracy of 78.4 and 80.1, respectively. In contrast, the best-performing baseline (PROCAP) achieves a maximum accuracy and macro F1 score of 75.10 and 74.85 on **FHM**, and 73.63 and 73.42 on **MAMI**.

Tag generation module: As stated earlier, Table 4 presents the tag generation results over **TOXICTAGS** dataset as well. The metrics used are CHRF, METEOR and SEMANTIC-SIMILARITY[19]. For each ground-truth tag, we compute each of the above metrics with all the generated tags and take the maximum value among these. Next, we average this maximum over all the ground-truth

[19] https://huggingface.co/tasks/sentence-similarity

Model	FHM		MAMI	
	acc	mF1	acc	mF1
PH	72.98	72.24	70.31	70.18
PC	75.10	74.85	73.63	73.42
MH	57.60	53.88	69.05	68.78
BM	66.00	65.80	70.50	70.10
StemTox	**78.4**	**78.11**	**80.1**	**79.67**

Table 5: Performance comparison of our method with baseline approaches on the widely used benchmark datasets FHM and MAMI. Best results are **highlighted.** Our model outperforms baseline methods in terms of accuracy (Acc) and macro-F1 (mF1) across both datasets. PH: PromptHate, PC: ProCap, MH: ModHate, BM: Bridging Modalities.

model	tag similarity		
	chrF	met	SS
RAM++	0.1942	0.058	0.406
RAM++ (O)	0.094	0.012	0.279
RAM	0.19	0.05	0.40
RAM (P)	0.087	0.014	0.338
RAM (O)	0.079	0.013	0.285
Tag2Text	0.1631	0.041	0.372
Tag2Text (P)	0.2102	0.087	0.332
StemTox	**0.4872**	**0.265**	**0.626**

Table 6: Performance of the tag generation module. Best results are **highlighted.** chrF: chrF, met: Meteor, SS: Semantic-Similarity, (O): open-set, (P): pre-trained using **ToxicTags**.

tags and calculate mean of these averages over all the test data points. Among the two models used within the **StemTox** multi-tasking framework, PaliGemma consistently outperforms LLaVA in terms of chrF, Meteor and Semantic-Similarity. Specifically, PaliGemma achieves scores of 0.4872 (chrF), 0.265 (chrF) and 0.626 (Semantic-Similarity), whereas LLaVA attains 0.2648, 0.107 and 0.447, respectively. These results indicate that PaliGemma is more effective than LLaVA for tag generation in the multi-tasking framework **StemTox**.

Tag generation baselines: Since **StemTox** generate tags along with toxicity labels, it is imperative to compare its tag generation performance with state-of-the-art generators. To this purpose, we use three state-of-the-art baselines for evaluation namely Tag2Text (Huang et al., 2024), RAM (Zhang et al., 2024), and RAM++ (Huang et al., 2025). To ensure fair evaluation, we further pre-trained the Tag2Text and RAM models on our **ToxicTags** dataset, allowing them to suitably adapt to domain-specific toxic vocabulary. Table 6 shows that **StemTox** by far outperforms all the other tag generation baselines in terms of all the three metrics – chrF, Meteor and Semantic-Similarity. State-of-the-art models fail to capture the meme's social context and their implicit meaning. They focus primarily on the visible objects in the image (refer to Table 7; also refer to Table 9 in Appendix for more examples).

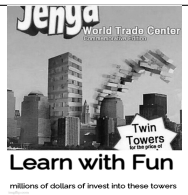
Image		
GT	trainers, hitler	911 9/11 twin towers impact, world trade center, jenga, 9/11 truth movement
ST	hitler, trainers, nazi	911, jenga, twin towers
R++	blue, man, shoe , smile	building, city
R++(O)	Athletic shoe, Close-up, Military person, Outdoor shoe	Close-up, Toy block
R	blue , man , shoe	building, city
R(P)	–	–
R(O)	Athletic shoe, Black-and-white, Military person, Outdoor shoe	Toy block
T2T	shoe, picture, uniform, person, man\nUser Specified	building, city, game\nUser Specified
T2T(P)	death, hitler, adolf hitler, holocaust, family, shoes\nUser Specified	9/11, twin towers, 911 9/11 twin towers impact\nUser Specified

Table 7: Examples of tags generated by different models. GT:Ground Truth, ST: **StemTox**, R: RAM, R++: RAM++, R(O): RAM open-set, R++(O): RAM++ open-set, T2T: Tag2Text, R(P): RAM pre-trained, T2T(P): Tag2Text pre-trained using **ToxicTags**. Please find more examples in Table 9 in the Appendix.

8 Error analysis

To understand the failures of the model we conduct an error analysis that helps to identify how these models struggle in classification; due to systematic weaknesses in contextual reasoning. We perform two complementary analyses to characterize these errors.

(i) We identify the most frequent tags for which each model struggles to predict the correct classification labels. **Key observations:** For LLaVA, the majority of the misclassifications are concentrated around tags corresponding to *death-related events* (`death, funeral, suicide, murder, heart attack`), *mental health and*

category	PALIGEMMA	LLAVA
hateful	0.2288 ± 0.2088	0.3029 ± 0.1968
dangerous	0.2423 ± 0.2043	0.3132 ± 0.1790
normal	0.3221 ± 0.1670	0.3478 ± 0.1557
offensive	$\mathbf{0.4950 \pm 0.1210}$	$\mathbf{0.5559 \pm 0.1065}$

Table 8: Average uncertainty scores ($\pm$ standard deviation) across different toxicity categories for PALIGEMMA and LLAVA. Higher numbers signify higher uncertainty in the model's classification.

violence (`depression, gun control, slavery`), and *religious or cultural entities* (`religion, Jesus, Bible`), indicating a bias toward emotionally sensitive scenes. In the case of PALIGEMMA, higher error frequencies are observed for tags associated with *humor and internet slang* (e.g., `comic, lol, oof size large, cursed`), suggesting that the model is unable to effectively interpret implicit sarcasm and meme culture.

(ii) To better understand where the models fail, we group all false positive samples according to their incorrectly predicted classes and perform an uncertainty-based error analysis. We quantify the model's behavior using an uncertainty score derived from the Maximum Softmax Probability (MSP) (Hendrycks and Gimpel, 2017). This uncertainty score is defined as:

$$\mathcal{U}(x) = 1 - \max_{c} p_\theta(c \mid x) \tag{7}$$

where x denotes a multimodal input sample, $c \in \{\textit{hateful}, \textit{offensive}, \textit{dangerous}, \textit{normal}\}$ represents a toxicity class, and $p_\theta(c \mid x)$ is the model-predicted posterior probability given input x.

Key observations: Table 8 shows that LLAVA consistently exhibits higher uncertainty for false positive predictions. Conversely, PALIGEMMA shows lower confidence margins. Both models display the highest uncertainty in the *offensive* category, indicating difficulty in learning accurate patterns, whereas interestingly the lowest uncertainty is observed in the *hateful* category.

9 Conclusion

This work makes several key contributions toward advancing research in toxic meme detection and multimodal content moderation. **First**, we curate a richly annotated dataset of 6,300 real-world meme-based posts through a two-stage labelling process which we call **TOXICTAGS**. **Second**, we enhance contextual understanding by introducing auxiliary metadata including meme titles and, most importantly, collaborative tags. **Third**, we propose **STEMTOX** that detects toxicity as well as generates tags given an arbitrary input meme. **Finally**, we compare the performance of toxicity detection and tag generation on an array of benchmarks and baselines for generalizability. Collectively, our contributions address key bottlenecks in semantically aligned toxic meme moderation and thus provides a foundation for building more accurate, context-aware, and socially responsible systems.

10 Limitations

While this work offers several important contributions, it also has a few limitations. ***First***, our dataset is limited to image-based memes, excluding other emerging modalities such as video and audio, which are increasingly used to disseminate harmful content. Future work will aim to extend this framework to multimodal datasets that better reflect current online communication trends. ***Second***, we do not examine the psychological and emotional impact of hateful memes on viewers. Exposure to such content may contribute to anxiety, depression, or desensitization – an important area that lies beyond the current scope. Addressing this limitation would require collaboration with psychologists, mental health experts, and affected communities. ***Third***, we do not investigate the real-world consequences of online hate memes, particularly their potential to incite offline violence or criminal behaviour. Understanding this link between online toxicity and offline harm is a critical direction for future research.

Acknowledgments

This work was supported in part by the **Scheme for Promotion of Academic and Research Collaboration (SPARC)**, Ministry of Education, Government of India, and by the **Anusandhan National Research Foundation (ANRF)** through a Core Research Grant (CRG). We gratefully acknowledge their support.

References

Carlos Arcila Calderón, Patricia Sánchez Holgado, Jesús Gómez, Marcos Barbosa, Haodong Qi, Alberto Matilla, Pilar Amado, Alejandro Guzmán, Daniel López-Matías, and Tomás Fernández-Villazala. 2024. From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and LGBT communities. *Humanit. Soc. Sci. Commun.*, 11(1).

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.

Fariha Tahosin Boishakhi, Ponkoj Chandra Shill, and Md. Golam Rabiul Alam. 2021. Multimodal hate speech detection using machine learning. In *2021 IEEE International Conference on Big Data (Big Data)*, page 4496–4499. IEEE.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. Pro-cap: Leveraging a frozen vision-language model for hateful meme detection. In *Proceedings of the 31st ACM international conference on multimedia*, pages 5244–5252.

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

Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. 2024a. Modularized networks for few-shot hateful meme detection. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4575–4584, New York, NY, USA. Association for Computing Machinery.

Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. 2024b. Modularized networks for few-shot hateful meme detection.

Guanqi Chen, Lei Yang, Guanhua Chen, and Jia Pan. 2023. Evaluating explanation methods for vision-and-language navigation.

Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. 2025. Multimodal language models see better when they look shallower. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6677–6695, Suzhou, China. Association for Computational Linguistics.

Yi Dai, Hao Lang, Kaisheng Zeng, Fei Huang, and Yongbin Li. 2023. Exploring large language models for multi-modal out-of-distribution detection. *arXiv preprint arXiv:2310.08027*.

Mithun Das and Animesh Mukherjee. 2023a. Banglaabusememe: A dataset for bengali abusive meme classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15498–15512.

Mithun Das and Animesh Mukherjee. 2023b. Transfer learning for multilingual abusive meme detection. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 245–250.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

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.

Zoe Wanying He, Sean Trott, and Meenakshi Khosla. 2025. Seeing through words, speaking through pixels: Deep representational alignment between vision and language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35645–35660.

Ming Shan Hee, Aditi Kumaresan, and Roy Ka-Wei Lee. 2024. Bridging modalities: Enhancing cross-modality hate speech detection with few-shot in-context learning. *arXiv preprint arXiv:2410.05600*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. 2025. Open-set image tagging with multi-grained text supervision. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 4117–4126, New York, NY, USA. Association for Computing Machinery.

Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. 2024. Tag2text: Guiding vision-language model via image tagging. In *ICLR*.

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.

Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163.

Girish A. Koushik, Helen Treharne, and Diptesh Kanojia. 2026. Tandem: Temporal-aware neural detection for multimodal hate speech.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, Asif Ekbal, and Vinutha B. NarayanaMurthy. 2024. CM-off-meme: Code-mixed Hindi-English offensive meme detection with multi-task learning by leveraging contextual knowledge. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3380–3393, Torino, Italia. ELRA and ICCL.

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2026. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *ACM Trans. Intell. Syst. Technol.*, 17(4).

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

Junyu Lu, Bo Xu, Xiaokun Zhang, Haohao Zhu, Kaichun Wang, Liang Yang, and Hongfei Lin. 2025. 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*, pages 559–569.

Grace Luo, Trevor Darrell, and Amir Bar. 2024. Vision-language models create cross-modal task representations. *arXiv preprint arXiv:2410.22330*.

Atanu Mandal, Gargi Roy, Amit Barman, Indranil Dutta, and Sudip Kumar Naskar. 2024. Attentive fusion: A transformer-based approach to multimodal hate speech detection.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

OpenAI. 2024. GPT-4o — openai.com. `https://openai.com/index/hello-gpt-4o/`.

Nirmalendu Prakash, Han Wang, Nguyen Khoi Hoang, Ming Shan Hee, and Roy Ka-Wei Lee. 2023. Promptmtopic: Unsupervised multimodal topic modeling of memes using large language models. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 621–631, New York, NY, USA. Association for Computing Machinery.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar,

Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796.

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

Yihan Ma1 Xinyue Shen1 Yiting Qu, Ning Yu2 Michael Backes, and Savvas Zannettou3 Yang Zhang. 2025. From meme to threat: On the hateful meme understanding and induced hateful content generation in open-source vision language models. In *USENIX Security Symposium (USENIX Security). USENIX*.

Naquee Rizwan, Paramananda Bhaskar, Mithun Das, Swadhin Satyaprakash Majhi, Punyajoy Saha, and Animesh Mukherjee. 2025a. Exploring the limits of zero shot vision language models for hate meme detection: The vulnerabilities and their interpretations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1669–1689.

Naquee Rizwan, Nayandeep Deb, Sarthak Roy, Vishwajeet Singh Solanki, Kiran Garimella, and Animesh Mukherjee. 2025b. Toxicity begets toxicity: Unraveling conversational chains in political podcasts. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 11776–11784, New York, NY, USA. Association for Computing Machinery.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing llms for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.

Surendrabikram Thapa, Farhan Ahmad Jafri, Kritesh Rauniyar, Mehwish Nasim, and Usman Naseem. 2024. Ruhate-mm: Identification of hate speech and targets using multimodal data from russia-ukraine crisis. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1854–1863.

Surendrabikram Thapa, Aditya Shah, Farhan Ahmad Jafri, Usman Naseem, and Imran Razzak. 2022. A multi-modal dataset for hate speech detection on social media: Case-study of russia-ukraine conflict. In *CASE 2022-5th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, Proceedings of the Workshop*. Association for Computational Linguistics.

Sriram Vajapeyam. 2014. Understanding shannon's entropy metric for information.

Wenxuan Wang, Xiaoyuan Liu, Kuiyi Gao, Jentse Huang, Youliang Yuan, Pinjia He, Shuai Wang, and Zhaopeng Tu. 2025. Can't see the forest for the trees: Benchmarking multimodal safety awareness for multimodal LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16993–17006, Vienna, Austria. Association for Computational Linguistics.

Qiyao Xue, Yuchen Dou, Zheyuan Ryan Shi, Xiang Lorraine Li, and Wei Gao. 2026. Mmbert: Scaled mixture-of-experts multimodal bert for robust chinese hate speech detection under cloaking perturbations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):34196–34204.

Akhila Yerukola, Saadia Gabriel, Nanyun Peng, and Maarten Sap. 2025. Mind the gesture:

Evaluating ai sensitivity to culturally offensive non-verbal gestures.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. 2024. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732.

A Ethics statement

We strictly adhered to the policies of the social media platforms from which the dataset was curated. All memes and associated metadata were manually collected from publicly available content after agreeing to relevant copyright terms, and user anonymity was preserved throughout the process. During annotation, we followed detailed guidelines to safeguard annotators' mental well-being and conducted two-stage annotation with pilot studies to reduce subjective bias. We acknowledge the potential for bias in both annotation and model predictions; to mitigate this, we employed diverse evaluation metrics and experimental setups to assess model robustness. This study complies with ethical standards for research involving publicly available data, with a focus on transparency, privacy, and minimising harm. Our work aims to support the broader effort to combat online hate while remaining mindful of its own limitations.

B Prompt templates

We display below the prompts used for experiments.

Prompts used for fine-tuning.

Generate tags by considering image, ocr and title

System prompt for zero-shot and few-shot classification tasks with VLMs.

You are an AI assistant tasked with classifying memes.
Consider the following definitions.
- hateful: {definition of hatful}
- dangerous: {definition of dangerous}
- offensive: {definition of offensive}
- normal: {definition of normal}

Based on the above definitions, the input image, title and the extracted OCR text from the image delimited by three backticks classify the meme as hateful, dangerous, offensive or normal.

Example output for hateful meme: {hateful.}
Example output for dangerous meme: {dangerous.}
Example output for offensive meme: {offensive.}
Example output for normal meme: {normal}

System prompt for zero-shot classification tasks with LLMS.

You are an AI assistant tasked with classifying memes.
Consider the following definitions.
- hateful: {definition of hatful}
- dangerous: {definition of dangerous}
- offensive: {definition of offensive}
- normal: {definition of normal}

Based on the above definitions and associated tags delimited by three backticks, classify the meme as hateful, dangerous, offensive or normal.

Example output for hateful meme: {hateful.}
Example output for dangerous meme: {dangerous.}
Example output for offensive meme: {offensive.}
Example output for normal meme: {normal}

C Platforms

C.1 Data curation platform

As stated earlier, we use https://imgflip.com as the curation platform for our data since it is centred on conversations using memes. We specifically use the *streams*[20] feature of the platform to collect these memes. The selected streams and their descriptions are outlined as follows:

DARK_HUMOUR (∼11k followers)[21] – *Welcome to Dark_humour, Imgflip's premier community for offensive humour. Stream mood: Relax liberals, it's called dark humour.*

MEMES_OVERLOAD (∼9.5k followers)[22] – *Hey there, welcome to MEMES_OVERLOAD, Imgflip's 4th largest stream for memes! We're all here to have fun, so make sure to follow the rules, and keep on memeing on.*

POLITICS (∼5k followers)[23] – *Humor and discussion around U.S. and world politics. Criticisms*

[20] https://imgflip.com/streams
[21] https://imgflip.com/m/Dark_humour
[22] https://imgflip.com/m/MEMES_OVERLOAD
[23] https://imgflip.com/m/politics

and debates are encouraged, but be constructive and don't harass anyone.

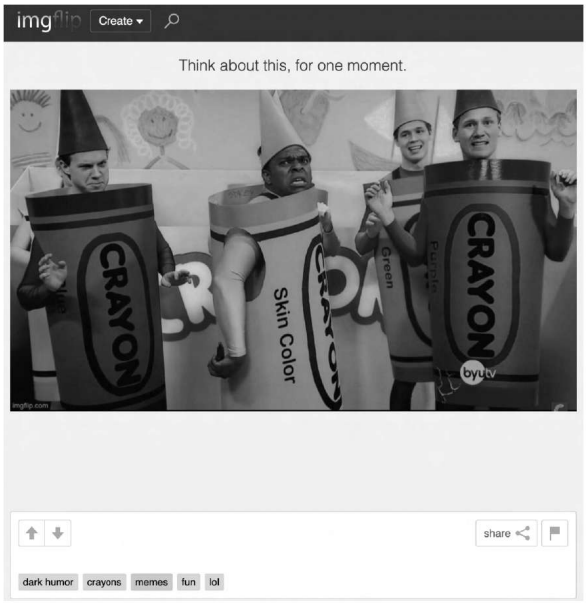


Figure 4: Post containing meme, title and tags from imgflip platform.

An example cropped post containing the corresponding title and tags is presented in Figure 4.

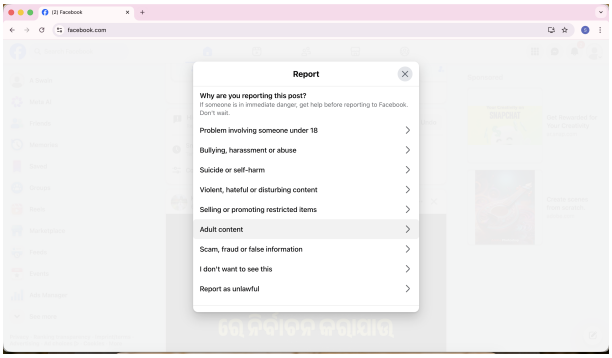
D Details of annotation

Figure 2 presents a brief pipeline of the employed annotation process over two stages. The instructions that we designed are given below, and some samples from those provided to annotators are present in Table 10. During the whole annotation process, annotators were asked to adhere to the definitions provided in the subsection D.1.

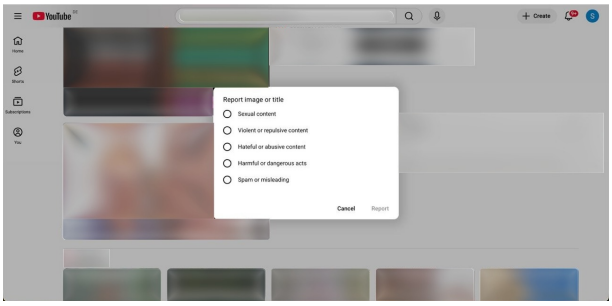
A– Each annotator was provided with a separate PDF containing the meme, title, tags and the OCR extracted text from the meme. Along with that, a separate Google sheet was also provided to reduce annotation bias and ensure fairness.

B– In Stage I, they were asked to strictly adhere to the provided definition of toxicity and segregate the memes as *toxic* or *normal*. Wherever confusion arose, annotators discussed in our daily scrum and through our Slack workspace.

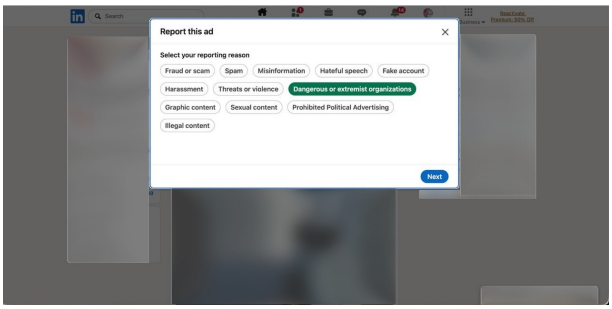
C– In Stage II, we provided the annotators with



(a) Facebook



(b) YouTube



(c) LinkedIn

Figure 5: Snapshots of reporting or policy interfaces from major social media platforms.

toxic memes based on majority voting as per Stage I. Note that before starting Stage II they were fairly aware of the type of content moderation required on these social media platforms. As a first step, they were asked to segregate hateful content, then from the remaining to identify dangerous samples; finally left with offensive samples, which were also verified. All the annotations were performed by strictly adhering to the definition of *hateful*, *dangerous* and *offensive* labels.

D.1 Definitions

hateful – *Reference*: (Kiela et al., 2020) – A direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease. 'At-

Image		
ground truth	alabama, incest, baby	chocolate, rivers, africa
STEMTOX	alabama, problems, abortion, pregnancy	chocolate, rivers, africa
RAM++	anime, brush, girl, person, kiss, love	boy, trench, floor, person, land, lay, man, mud, puddle, water
RAM++ (O)	Close-up	Jeans
RAM	anime, girl, person, kiss	boy, floor, person, land, lay, man, mud, puddle, water
RAM (P)	–	–
RAM (O)	–	Ribs
TAG2TEXT	anime, girl, people, person User Specified	mud, puddle, water, man User Specified
TAG2TEXT (P)	women, psychopath\nUser Specified	repost, christmas\nUser Specified

Table 9: Comparative examples of tags generated by different models. (O): open-set, (P): pre-trained using **TOXICTAGS**.

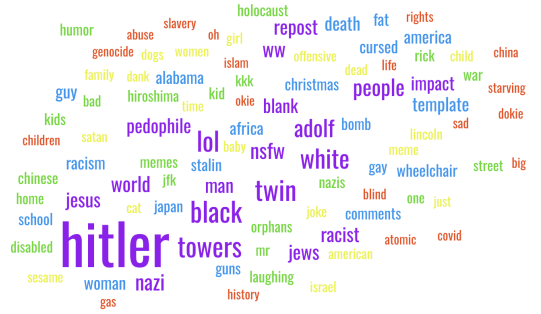
Image				
Title	Does Anyone Else See The Problem With This Advertising Sign?	Cannibal cafe menu	If ykyk	My friend sent me this
Tags	signs, advertising	hannibal lecter, cannibals, cannibal, cannibalism	if you know you know, porn	if you know you know, the owl house, bee movie
OCR extracted text	imaflip.com ginger === S	CANNIBAL CAFE MENU BABY BACK RIBS FINGER SANDWICH OPEN FACE SANDWICH STU STEW TOE FU SCRAMBLED LEGS BAKED FRIARS MAC CHEESE (VEGAN)	Consider John Frazzled FrazzleMyGimp PIZZA GUY: Your total is $26.34 ME: I can't afford that PIZZA GUY: Pay another way PORN DIRECTOR: Cut	According to some, this will confuse children But this is completely fine HOME TEETH 3
Expert annotation	**hateful**	**dangerous**	**offensive**	**normal**

Table 10: Four memes from the expert annotation samples provided to the annotators.

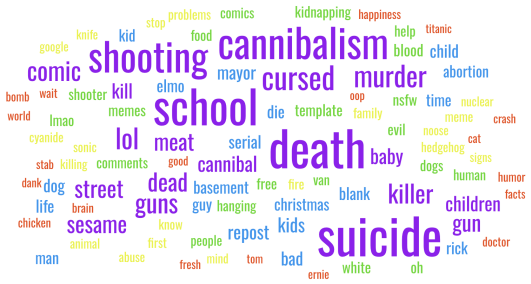
tack' is defined as violent or dehumanising (comparing people to non-human things, e.g., animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking hate crime is also considered hateful.

dangerous – *Reference*[24] – A text, meme or speech which is not hateful but uses any form of expression that can increase the risk of its audience to condone or participate in violence against members of another group will be considered dan-

[24] https://www.dangerousspeech.org/dangerous-speech

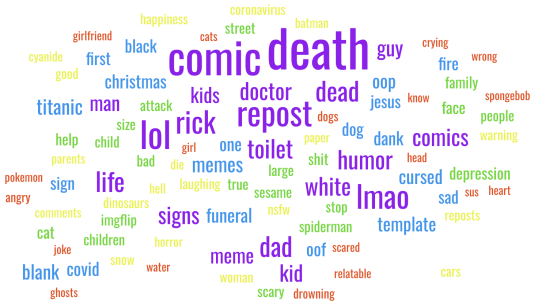


hateful



dangerous

offensive



normal

Table 11: Label-wise word cloud of tags.

gerous.

offensive – *References*: (Roy et al., 2023; Mathew et al., 2021) – A text, meme or speech which is neither hateful nor dangerous but uses abusive slurs or derogatory terms will be considered offensive.

toxic – *Reference*: PerspectiveAPI[25] – A rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.

normal – A meme which is not toxic and follows social norms.

D.2 Safety guidelines

Primarily, we performed the following steps to keep our annotators mentally safe with such content:

(i) Only 50 samples per day were provided to them, wherein we updated the PDF and added corresponding samples in the Google sheet.

(ii) We conducted 15 minutes of daily mental well-being sessions in our daily scrum by adopting various online activities suggested by WHO. [26]

(iii) We also asked the annotators to agree to the data source platform's terms and policies. Further, we strictly ensured that they did not disclose the identity of users in the provided Google sheet.

[25] Perspective API Model Card

[26] https://www.who.int/news-room/feature-stories/mental-well-being-resources-for-the-public

E Further analysis of the dataset

Word clusters of tags: Figure 11 presents the word cloud of tags for *hateful*, *dangerous*, *offensive* and *normal* memes. We can clearly segregate the word clouds for hateful and dangerous, hence signifying our initial observations. For offensive and normal, even though they have similar words, the presence of terms like *cursed, nsfw* and usage of comic character terms like *sesame, street* significantly separates them.

F Employed models

PaliGemma: We use PaliGemma-2 which presents a series of models of varying size with further improved capabilities due to the incorporation of Gemma-2 LLM instead of Gemma. We use `google/paligemma2-10b-pt-224` version from HuggingFace in this work.

LLaVA: We use LLaVA-1.5 (Liu et al., 2024) that has shown significant improvement over its prior models. We use `llava-hf/llava-1.5-7b-hf` checkpoint for running our experiments.

GPT-4o: One of the first models by OpenAI team to have multimodal capability, GPT-4o (OpenAI, 2024) has proved its wide applicability across

Image			
Title Gotta go fast	All members get 69% off all items	Rouge.exe is getting desperate!	
Tags	gotta go fast, sonic the hedgehog, **abortion**	**human stupidity**, ernie and bert, **traffic light**	rougeexe, sonic the hedgehog, **free candy van**
OCR extracted text	Just get the back alley abortion and learn to keep your legs closed! Welp. Time to go buy milk. Fast. But sonic, it's YOUR baby!	Ernie and Bert want to create a human trafficking program. Join today and help make a difference in your community!	FREE BAT FLAVORED SOUP! JUST GET IN THE $#%&ING VAN. IT ONLY HURTS TIL YOU PASS OUT! MASKS SAVE LIVES
Ground truth annotation	**dangerous**	**dangerous**	**dangerous**

Table 12: Dangerous memes: portrayal of danger through anime/cartoon characters.

multiple domains[27].

MISTRAL: MISTRAL is an instruction-tuned large language model designed for efficient reasoning and strong performance across diverse NLP tasks. We use `mistralai/Mistral-7B-Instruct-v0.3` checkpoint from Huggingface.

LLAMA: We use LLAMA-3.1is which is a part of Meta's LLaMA series, designed for high-quality and robust generalization. We use `meta-llama/Llama-3.1-8B-Instruct`.

License agreement: We agreed to the terms of usage of all employed VLMs before using them in our work.

G Experimental setup

We train both the uni- and multi-tasking framework using a consistent experimental configuration to ensure a fair comparison. For the multi-task setting, the training objective consists of two components: a weighted cross-entropy loss for the classification task and the default autoregressive generation loss for toxic tag generation. In the case of uni-task learning, there is no generation loss. For computing class weights, we use the following weighting scheme: . All input sequences are truncated or padded to a fixed maximum sequence length of 2048 tokens, and the random seed is fixed to 42 to ensure reproducibility.

Training is performed for three epochs with a per-device batch size of 1 and a gradient accumulation step of 4, resulting in an effective batch size of 4. We use the `AdamW` optimizer with a learning rate of 2×10^{-5}, weight decay of 1×10^{-6}, and β_2 set to 0.999, along with a warm-up of two steps. For memory-efficient training, the model is loaded using 8-bit quantization via the `BitsAndBytes` configuration, with computations carried out in `FP16` precision. For the multi-task Learning framework, we apply LoRA-based parameter-efficient fine-tuning with rank 8, targeting the following modules: {`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`}.

In contrast, for the uni-task learning framework, the model backbone is fully frozen during training. Extracted hidden representations are normalized using the default normalization layers provided by the backbone model at the time of forward propagation. A sequence compression module is used to extract features into a fixed-dimensional embedding, which is then passed through a label classification head to predict the final class labels.

For the **TOXICTAGS** dataset, we fix the number of classes to four while for the **FHM** and **MAMI** datasets, we fix the number of classes to two. For result calculation and analysis of stage I, we merge the hateful, dangerous, and offensive categories into a single toxic class. This aggregation is the reverse of the two-stage annotation strategy followed during dataset construction.

[27] https://learn.microsoft.com/en-us/azure/ai-services/openai/