MemeGuard: An LLM and VLM-based Framework for Advancing Content Moderation via Meme Intervention

Prince Jha[1], **Raghav Jain**[1], **Konika Mandal**[1], **Aman Chadha**[2*], **Sriparna Saha**[1], and **Pushpak Bhattacharyya**[3]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna
[2]Amazon AI
[3]Department of Computer Science and Engineering, Indian Institute of Technology Bombay

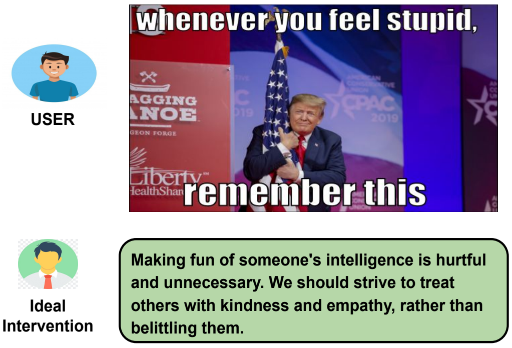
Abstract

In the digital world, memes present a unique challenge for content moderation due to their potential to spread harmful content. Although detection methods have improved, proactive solutions such as intervention are still limited, with current research focusing mostly on text-based content, neglecting the widespread influence of multimodal content like memes. Addressing this gap, we present *MemeGuard*, a comprehensive framework leveraging Large Language Models (LLMs) and Visual Language Models (VLMs) for meme intervention. *MemeGuard* harnesses a specially fine-tuned VLM, *VLMeme*, for meme interpretation, and a multimodal knowledge selection and ranking mechanism (*MKS*) for distilling relevant knowledge. This knowledge is then employed by a general-purpose LLM to generate contextually appropriate interventions. Another key contribution of this work is the ***I**ntervening **C**yberbullying in **M**ultimodal **M**emes (ICMM)* dataset, a high-quality, labeled dataset featuring toxic memes and their corresponding human-annotated interventions. We leverage *ICMM* to test *MemeGuard*, demonstrating its proficiency in generating relevant and effective responses to toxic memes.[1]

Disclaimer: *This paper contains harmful content that may be disturbing to some readers.*

1 Introduction

In today's digital world, memes serve as a universal language for expression and engagement. However, as they become a powerful tool for rapid information dissemination, they are increasingly weaponized for cyberbullying and spreading toxic content, posing a challenge to existing content moderation systems. These systems struggle to decipher memes' nuanced meanings, typically reacting (Ohlheiser, 2016) rather than proactively mitigating the harm of offensive content.



Making fun of someone's intelligence is hurtful and unnecessary. We should strive to treat others with kindness and empathy, rather than belittling them.

Figure 1: An instance of the meme intervention task.

Intervention represents a proactive approach to content moderation, going beyond simple detection to take preventive action against offensive content. Interventions aim to mitigate the harmful effects of toxic content and foster a more positive and respectful online discourse. However, existing intervention research has primarily been limited to text-based content such as hate speech (Qian et al., 2019), and misinformation (He et al., 2023). The excessive focus on text-based content overlooks the prevalence of multimodal content, which are major contributors to the content ecosystem in social media platforms. This exacerbates the potential for multimodal toxicity, enabling misuse of such mediums. *A representative example of intervening in case of a cyberbullying meme is displayed in Figure 1.*

Large Language Models (LLMs) (Kojima et al., 2023), and Visual Language Models (VLMs) (Ghosh et al., 2024b) have shown remarkable capability in understanding and generating human-like text and multimedia content. This has led to their application in various tasks within the domain of content moderation, but predominantly for detection purposes. Their ability to understand nuanced language and visual cues has been leveraged to detect toxic or harmful content, both in text (ElSherief et al., 2021; Maity et al., 2023, 2024) and multimodal formats (Maity et al., 2022; Jain

[*] Work does not relate to position at Amazon.

[1] Code and dataset are available at https://github.com/Jhaprince/MemeGuard

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8084–8104
August 11-16, 2024 ©2024 Association for Computational Linguistics

et al., 2023; Jha et al., 2024). Some attempts have also been made to use these models for intervention generation tasks, but these efforts have been largely restricted to text-based content (Qian et al., 2019; He et al., 2023). In this landscape, the zero-shot learning (Dong et al., 2023) of these models presents a unique advantage, particularly for tasks like meme intervention, which are characterized by data scarcity.

While these models hold considerable promise for content moderation, they are not without their limitations. First and foremost, vanilla LLMs and VLMs lack grounding in a knowledge base specific to the task at hand. This can lead to the generation of generic interventions that may fail to adequately address the bias, stereotype, assertions, and toxicity present in the meme content. In terms of visual content, while VLMs have performed well on a variety of traditional visual-linguistic tasks, they often struggle when it comes to memes. The primary reason behind this is the unique nature of memes, which are highly contextual and often rely on a shared understanding of internet subcultures for their interpretation. This makes the accurate assimilation and analysis of memes a challenging task. Finally, even when these models are grounded in a knowledge base for the task of meme intervention, there is a need to filter irrelevant and noisy information. Without appropriate filtering, these models might end up incorporating irrelevant or misleading information into their interventions.

In order to address these limitations inherent in LLMs and VLMs for meme intervention, we developed a comprehensive framework called ***MemeGuard***. The development of *MemeGuard* was a multi-stage process designed to create a tool capable of understanding and effectively intervening in the spread of toxic memes. In the first stage, we developed a meme-aligned VLM (***VLMeme***), specifically fine-tuned to understand and interpret memes in all their complexity. This allowed our model to delve deeper into the content of memes. Next, we utilized this meme-aligned VLM to identify various facets of the meme, such as the underlying toxicity, bias, stereotypes, and claims being made, which provides valuable insights into the meme's potential harm. To address the challenge of irrelevant knowledge, we then proposed a Multimodal Knowledge Selection mechanism (***MKS***). This mechanism retained only the most relevant knowledge for the intervention generation process. In the subsequent stage, we utilized a general-purpose LLM grounded on this refined knowledge to generate appropriate interventions. This model took the insights provided by the VLM and the ranked knowledge to create contextually relevant and effective responses to toxic memes. Finally, to test the efficacy of our framework, we developed a high-quality labeled dataset featuring a variety of toxic and cyberbully memes with their corresponding human annotated intervention, ***I****ntervening* ***C****yberbullying in* ***M****ultimodal* ***M****emes (ICMM)* dataset. To summarize, we make the following main contributions:

A novel task of Meme Intervention to combat the toxicity of cyberbullying memes.

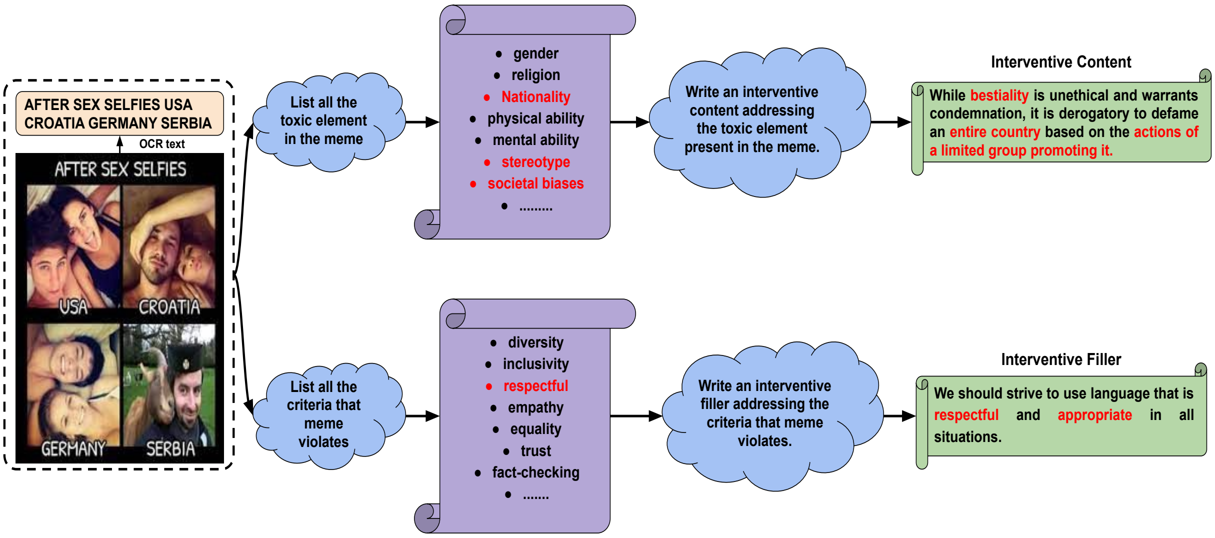
A novel dataset, *ICMM*, to advance the research in this area.

A novel framework, *MemeGuard*, that utilizes a meme-aligned VLM (*VLMeme*) to generate contextual information about the meme that is then used to generate the final intervention.

2 Related Works

Meme Analysis: Meme analysis, a fast-developing field, computationally scrutinizes memes, multimodal entities blending text and visuals, to detect harmful elements like hate speech, offensiveness, cyberbullying, and stereotypes. Kiela et al. (2020) proposed a benchmark dataset on hateful memes. Pramanick et al. (2021b) extended the HarMeme dataset (Pramanick et al., 2021a) with additional memes related to COVID-19 and US politics. Maity et al. (2022) created a cyberbullying meme dataset, which is the only publicly available meme dataset in code-mixed language. Furthermore, Mathias et al. (2021) extended Facebook's meme dataset (Kiela et al., 2020) to include two subtasks to identify the attacked category and determine the attack type in memes. Lately, there has been an increase in research dedicated to understanding the context of memes. Hee et al. (2023) addressed the research gap by proposing *HatReD* dataset annotated with the underlying hateful contextual reasons. Moreover, Sharma et al. (2023) proposed a new task that aims to identify evidence from a given context to explain the meme. Jha et al. (2024) proposed MultiBully-Ex a benchmark dataset for with multimodal rationales for code-mixed cyberbullying memes. Recently, Hwang and Shwartz (2023) also released a labeled dataset for the task of meme caption.

Intervention and Counterspeech Generation:


AFTER SEX SELFIES USA CROATIA GERMANY SERBIA
OCR text
AFTER SEX SELFIES
USA
CROATIA
GERMANY
SERBIA
List all the toxic element in the meme
gender
religion
Nationality
physical ability
mental ability
stereotype
societal biases
....
Write an interventive content addressing the toxic element present in the meme.
Interventive Content
List all the criteria that meme violates
diversity
inclusivity
respectful
empathy
equality
trust
fact-checking
....
Write an interventive filler addressing the criteria that meme violates.
Interventive Filler

While bestiality is unethical and warrants condemnation, it is derogatory to defame an entire country based on the actions of a limited group promoting it.

We should strive to use language that is respectful and appropriate in all situations.

Figure 2: Flowchart depicting the annotation guideline illustrated with a sample example.

Counterspeech can be regarded as a preferred remedy for combating hate speech as it simultaneously educates the perpetrators about the consequences of their actions and upholds the principles of free speech. Recent studies have shown the remarkable effectiveness of social intervention on social media like Twitter (Wright et al., 2017), and Facebook (Schieb and Preuss, 2016). Wright et al. (2017) studied the conversations on Twitter and found that some arguments between strangers lead to favorable changes in discourse. Mathew et al. (2018) released the first counterspeech dataset with 13,924 manually annotated YouTube comments, marked as counterspeech or not. Due to the limited scalability and generalizability of socially intervened counterspeech datasets, expert-annotated datasets were developed (Qian et al., 2019; Chung et al., 2019) for counterspeech in hate speech.

3 ICMM Dataset

To create *ICMM*, we utilize *MultiBully* dataset (Maity et al., 2022), which includes 3222 bully and 2632 nonbully memes. We selected this dataset because it is the only openly available meme dataset on cyberbullying in a code-mixed setting. Our annotation process focuses solely on bully memes to develop interventions against online cyberbullying.

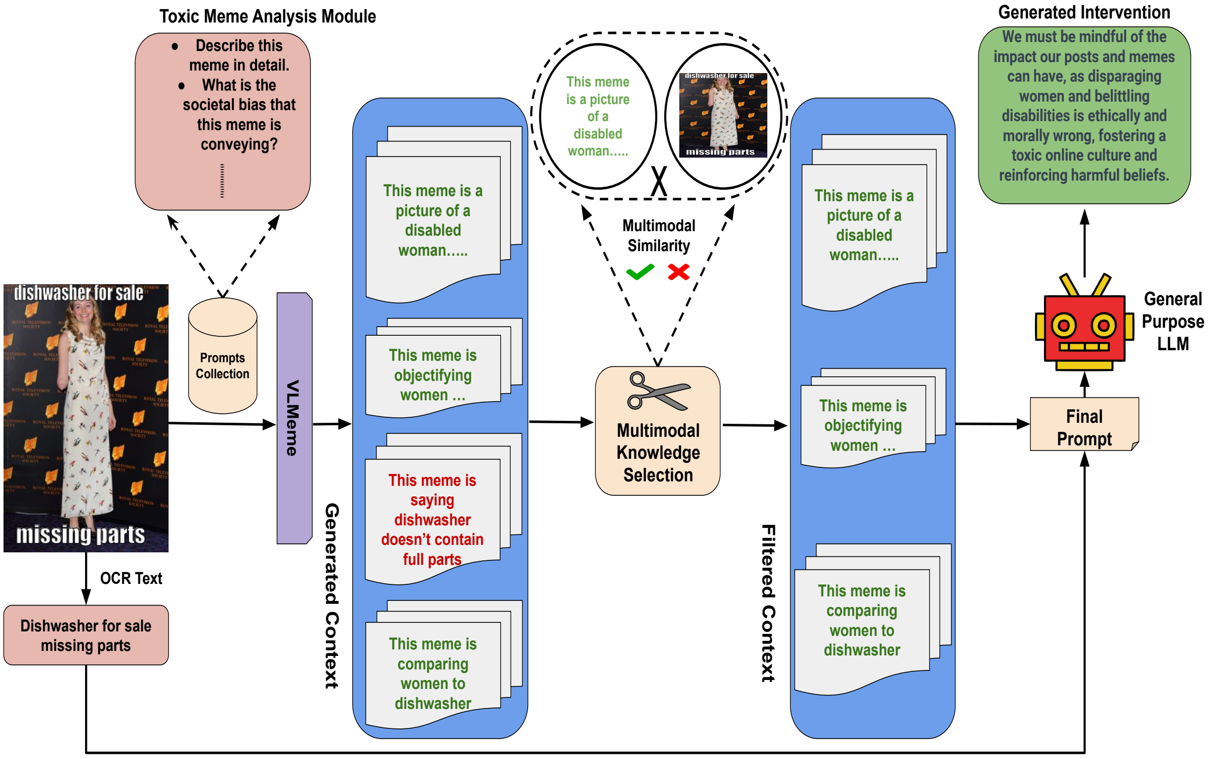
3.1 Annotation Training

The annotation team consisted of two expert annotators working in content moderation and three novice annotators, all computer science undergraduate students who possess proficiency in both Hindi and English. First, ten undergraduate computer science students were voluntarily hired through the department email list and compensated through honorarium. We provided novice annotators with 20 diverse expert-annotated interventions, along with a comprehensive set of annotation guidelines detailed in Appendix A.4 and illustrated in Figure 2 for reference. We have conducted four-phase training to ensure annotators were proficient and well-versed in the tasks. In each training phase, annotators are asked to write interventions for 20 cyberbullying memes. Later, expert annotators assessed the quality of interventions annotated by novice annotators based on fluency, adequacy, informativeness, and persuasiveness as mentioned in Appendix A.6. After the completion of each phase of training, expert annotators met with novice annotators to discuss how poorly rated intervention annotations could be improved. These discussions further trained annotators, and simultaneously, annotation guidelines were also renewed. As a result, the top three annotators were selected based on their performance across all phases, whose quality of intervention annotations significantly improved from the first phase (fluency = 3.97, adequacy = 2.59, informativeness = 2.44, and persuasiveness = 2.21) to the fourth phase (fluency = 4.91, adequacy = 4.81, informativeness = 4.79, and persuasiveness = 4.84).

3.2 Main Annotation

We used the open-source platform Docanno[2] deployed on a Heroku instance for main annotation (cf. Appendix A.7) where three qualified anno-

[2] https://github.com/doccano/doccano



Toxic Meme Analysis Module
Describe this meme in detail.
What is the societal bias that this meme is conveying?
Prompts Collection
dishwasher for sale
missing parts
OCR Text
Dishwasher for sale missing parts
VLMeme
Generated Context
This meme is a picture of a disabled woman.....
This meme is objectifying women ...
This meme is saying dishwasher doesn't contain full parts
This meme is comparing women to dishwasher
This meme is a picture of a disabled woman.....
Multimodal Similarity
✓ ✗
Multimodal Knowledge Selection
Filtered Context
This meme is a picture of a disabled woman.....
This meme is objectifying women ...
This meme is comparing women to dishwasher
Final Prompt
General Purpose LLM
Generated Intervention
We must be mindful of the impact our posts and memes can have, as disparaging women and belittling disabilities is ethically and morally wrong, fostering a toxic online culture and reinforcing harmful beliefs.

Figure 3: The proposed framework of our *MemeGuard* system. Sentences highlighted in green within the Generated Context block symbolize relevant knowledge, while those in red signify irrelevant knowledge.

tators were provided with secure accounts to annotate and track their progress exclusively. Each intervention annotation was assessed by two peer evaluators. The process of annotating interventions is both time-consuming and costly. On average, it takes approximately 6-8 minutes to write interventions for a single meme sample, ensuring high-quality annotation and capturing sufficient evidence and stereotypes for both interventive content and interventive filler and an additional 2-3 minutes for assessment. We offer an honorarium of 8 INR (Indian Rupee) per intervention sample annotation and 2 INR per evaluation, in line with India's minimum wage standards as outlined in the Minimum Wages Act, 1948 (mwa). As a result of these factors, we have decided to restrict the annotation to only 1000 test samples for this specific project, which is deemed sufficient for testing models based on existing literature. We have commenced annotations for five days a week, adhering to the schedule described in appendix A.5. It took approximately five weeks to complete the entire intervention annotation process. The interventive content has an average length of 14.71 while the interventive filler has an average length of 16.29. In the annotated interventions, we observed remarkably high average scores of 4.98, 4.87, 4.46, and 4.91 for fluency, adequacy, persuasiveness, and informativeness (cf. Appendix 3), respectively. Additionally, two evaluators achieved unanimous agreement scores of 94.7% for fluency ratings, i.e., evaluators rated the same score for 94.7% of time, 89.1% for adequacy ratings, 90.48% for informativeness ratings, and 82.39% for persuasiveness ratings.

4 Methodology

Problem Formulation: The input consists of two entries: (1) the source content M, which is a meme potentially bearing toxic content; this multimedia data is composed of visual elements inherent to the image; (2) the OCR text X derived from the source content M, representing the textual elements embedded within the meme. The output is a natural language sequence Y, which represents the intervention crafted to counter the toxic content of the meme M.

This section presents the novel architecture of ***MemeGuard*** (Figure 3), a multimodal, knowledge-grounded framework designed to mitigate toxic content in memes. For a comprehensive understanding, we partition our approach into three distinct modules: (i) **Toxic Meme Analysis Mod-**

ule (ii) **Multimodal Knowledge Selection (MKS)**, and (iii) **Intervention Generation Module**.

4.1 Toxic Meme Analysis Module

Development of VLMeme: Our model, ***VLMeme*** (Figure 4), is an enhancement built atop the MiniGPT-4 model (Zhu et al., 2023). MiniGPT-4 is designed to merge visual data from a proficiently trained vision encoder with the capabilities of an advanced large language model (LLM), thereby facilitating complex linguistic tasks. This model leverages the Vicuna language decoder (Chiang et al., 2023) – a model built atop LLaMA (Touvron et al., 2023) – in tandem with a visual encoder that employs a Vision Transformer (ViT) backbone (Dosovitskiy et al., 2021) and a pre-trained Q-Former (Zhang et al., 2023).

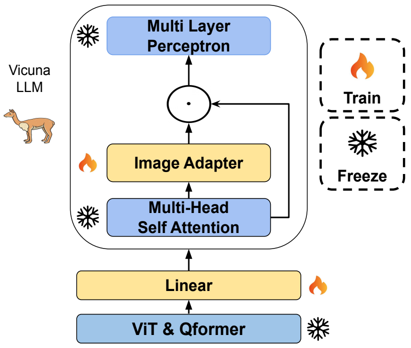


Figure 4: Architectural Diagram of *VLMeme*.

However, despite their effectiveness in handling traditional visual-linguistic tasks, Visual Language Models (VLMs) tend to face difficulties with meme content due to their inherent complexity and cultural connotations. To surmount this challenge, we augment the Vicuna language decoder with an image adapter inspired by recent works on efficient training of NLP models (Houlsby et al., 2019; Yuan et al., 2023), thereby improving its capacity to decipher visual information. Further, we fine-tune this enhanced model on a meme caption dataset (Hwang and Shwartz, 2023). This approach imbues our model with a deeper understanding of memes, facilitating its handling of the specific nuances and layered meanings that memes often embody. Formally, image adapter computation inside the Vicuna model is shown below:

$$Z_{IA} = Linear_1(RELU(Linear_2(Z_{MHA}))) + Z_{MHA} \quad (1)$$

where Z_{MHA} represents the output from Multi-Headed Attention layer of Vicuna, Z_{IA} represents the output from this image adapter layer, RELU represents the activation function computations, $Linear_1$, and $Linear_2$ represent the linear feed-forward neural network layers. Following the incorporation of the image adapter within the Vicuna language decoder, the subsequent step involves fine-tuning this enhanced model on a meme dataset. This dataset is composed of memes accompanied by their respective descriptions, serving as a rich resource for context-based meme understanding. For this purpose, we exploit the recently released *MEMECAP* (Hwang and Shwartz, 2023) dataset, which offers a wide range of memes with their associated descriptions. The fine-tuning strategy aligns with the method described in Zhu et al. (2023); Ghosh et al. (2024a). During this process, the components of Vicuna, Q-Former, and ViT are kept frozen, allowing us to focus specifically on the meme understanding abilities of our model. This will result in a meme-aligned vision language model ***VLMeme***.

Contextual Knowledge Generation: The next stage in the development of our framework involves utilizing ***VLMeme*** to generate contextual knowledge to gain a deeper understanding of the meme content, allowing it to identify the underlying toxicity, biases, stereotypes, and assertions present within the meme. Formally, for a given meme M, we formulate a set of prompts $P = \{p_0, p_1, \ldots, p_n\}$ with the aim of generating diverse contextual information about meme M, $KS = \{ks_0, ks_1, \ldots, ks_n\}$ by prompting ***VLMeme*** as follows: for each prompt p_i: $ks_i = \mathit{VLMeme}(p_i)$. Specifically, these prompts are designed to yield the following critical insights: (i) meme description, (ii) bias inherent in the meme, (iii) stereotypes propagated by the meme, (iv) toxic elements within the meme, and (v) assertions and claims conveyed through the meme. The detailed set of prompts is presented below.

Prompts used to generate contextual information

- Describe this meme in detail.
- What is the societal bias that this meme is conveying?
- What is the societal stereotype that this meme is conveying?
- What is the toxicity and hate that this meme is spreading?
- What are the claims that this meme is making?

4.2 Multimodal Knowledge Selection (MKS)

While we anticipate that ***VLMeme*** will generate relevant information, KS, there are instances where these models might deviate from the main query (Holtzman et al., 2020). This could introduce irrelevant context, distracting the Large Language Mod-

	Model	ROUGE			BLEU				Average Score		BERTScore
		R1	R2	RL	B1	B2	B3	B4	BLEUavg	Hmean	
Dolly	OCR	2.39	0.15	2.07	7.49	4.52	1.84	0.84	3.67	2.65	74.35
	OCR + MiniGPT4	0.94	0.05	0.74	2.77	1.73	0.75	0.36	1.40	0.97	73.81
	OCR + VLmeme	8.48	0.23	6.26	16.55	12.11	6.17	**3.36**	9.55	7.56	79.17
	MemeGuard	**8.68**	**0.41**	**7.22**	**17.05**	**12.88**	**6.65**	2.86	**9.86**	**8.33**	**82.27**
LLaMA	OCR	8.48	0.85	6.96	31.15	27.7	10.56	4.6	18.50	**10.11**	79.98
	OCR + MiniGPT4	8.84	0.87	**6.98**	31.19	27.77	10.64	4.74	18.58	10.15	79.98
	OCR + VLmeme	6.18	0.61	4.73	30.89	28.4	10.95	**4.78**	18.75	7.55	80.11
	MemeGuard	**9.27**	**1.71**	5.29	**31.96**	**29.08**	**11.89**	4.3	**19.31**	8.30	**80.11**
RedPajama	OCR	8.31	0.45	6.63	**39.17**	19.9	7.47	3.34	17.47	9.61	79.56
	OCR + MiniGPT4	**11.61**	0.7	8.55	33.94	22.19	10.6	5.2	17.98	11.58	82.4
	OCR + VLmeme	10.82	**0.58**	7.93	35.1	22.72	10.02	4.88	18.18	11.04	80.93
	MemeGuard	10.83	0.57	**9.01**	35.23	**22.82**	**11.05**	**4.9**	**18.5**	**12.11**	**83.42**
FLAN-T5	OCR	5.92	0.4	4.96	10.78	4.16	1.34	0.59	4.22	4.55	83.19
	OCR + MiniGPT4	13.88	1.03	10.13	39.65	26.68	12.75	6.58	21.41	13.75	84.5
	OCR + VLmeme	13.96	0.77	10.91	45.47	27.98	12.33	**7.22**	23.25	14.85	85.04
	MemeGuard	**14.11**	**2.01**	**11.22**	**48.91**	**30.21**	**12.81**	6.6	**24.63**	**15.41**	**87.21**
GPT3.5-Turbo	OCR	21.05	3.85	13.25	37.9	29.57	17.53	11.24	24.06	17.08	85.07
	OCR + MiniGPT4	13.71	1.83	9.3	21.3	17.99	10.97	6.51	14.19	11.23	85.33
	OCR + VLmeme	22.44	3.13	10.18	37.76	30.23	17.53	11.35	24.22	14.33	86.26
	MemeGuard	**24.42**	**4.7**	**15.22**	**39.45**	**32.32**	**17.82**	**11.91**	**25.37**	**19.03**	**89.02**

Table 1: Performances of various *MemeGuard* models and corresponding baselines, evaluated using automatic metrics with different base LLMs.

	Model	Fluency	Adequacy	Persuasiveness	Informativeness
FLAN-T5	OCR	4.27	2.69	2.16	2.72
	OCR + MiniGPT4	4.31	2.81	2.19	2.79
	OCR + VLMeme	4.35	3.26	2.23	3.19
	MemeGuard	4.39	3.32	2.24	3.27
GPT3.5-Turbo	OCR	4.79	3.6	3.32	3.57
	OCR + MiniGPT4	4.81	3.82	3.47	3.61
	OCR + VLMeme	4.81	3.95	3.91	4.05
	MemeGuard	4.82	4.16	3.97	4.13
Annotated Intervention		4.98	4.87	4.46	4.91

Table 2: Human evaluation scores of the two best *MemeGuard* models and their corresponding baselines across different metrics.

els (LLMs) (Shi et al., 2023), thereby hampering its ability to accurately analyze the meme's content and generate effective interventions. As such, there is a need to filter out this irrelevant context to maintain the effectiveness of the intervention process. To address this issue, we propose a filtering strategy named *Multimodal Knowledge Selection*, which works as follows. Formally, for each ks_i, which corresponds to a specific text field in the meme, it is further broken down into a set of m sentences $ks_i = \{s_1, s_2, \ldots, s_m\}$ using sentence tokenization[3]. To achieve this, we employ a pre-trained multimodal model, ImageBind (Girdhar et al., 2023), which serves as an off-the-shelf encoder-based multimodal model $enc(\cdot)$ that maps a token sequence (text) and an image to their respective feature vectors embedded in a unified vector space. Cosine similarity $sim(\cdot, \cdot)$ is then used to measure the relevance of each sentence to the image. Formally, a sentence s_j is retained if $sim(enc(s_j), enc(M)) > Th$, where M is the image associated with the meme, and Th is the predefined similarity threshold. Building upon this, we define the subset of retained knowledge sentences, having undergone the *MKS*, as follows:

$$ks'_i = \{s_j | s_j \in ks_i \text{ and } sim(enc(s_j), enc(M)) > Th\} \quad (2)$$

With this approach, the filtered knowledge sentences $KS' = \{ks'_0, \ldots, ks'_n\}$ not only hold a high degree of visual-textual alignment with the corresponding meme but are also contextually relevant.

4.3 Intervention Generation Module

This module is designed to utilize the refined knowledge sentences $KS' = \{ks'_0, ks'_1, \ldots, ks'_n\}$ to generate well-informed and contextually pertinent interventions for toxic memes. The Intervention Generation Module operates on a pre-trained Large Language Model (LLM), denoted as LM. To generate the intervention, the LM is prompted with a specially designed prompt P_{in}, which incorporates both the meme's OCR text X and the generated knowledge about the meme KS'. The output of this process, I, is the intervention text generated by the model. Formally, this process can be described as follows: $I = LM(P_{in}(X, KS'))$. The detailed prompt is presented below.

Prompt used to generate final intervention

> This is a toxic meme with the description: $\{ks'_0\}$. The following text is written inside the meme: $\{X\}$. Rationale: Bias: $\{ks'_1\}$, Toxicity: $\{ks'_2\}$, Claims: $\{ks'_3\}$, and Stereotypes: $\{ks'_4\}$. Write an intervention for this meme based on all this knowledge.

5 Experimental Results and Discussion

Experimental Setup: We leveraged following general purpose LLMs for intervention generation module: Dolly[4], LLaMA (Touvron et al., 2023), RedPajama[5], FLAN-T5 (Chung et al., 2022), and GPT3.5-Turbo. Details are mentioned in Appendix

[3] https://www.nltk.org/api/nltk.tokenize.html

[4] https://huggingface.co/databricks/dolly-v2-3b

[5] https://huggingface.co/togethercomputer/RedPajama-INCITE-Chat-3B-v1

Table 4. We tested these LLMs on different settings: (1) With only OCR text in prompt (OCR), With the knowledge obtained from MiniGPT4 (OCR+MiniGPT4), With the knowledge obtained from VLMeme (OCR+VLMeme), and then our proposed framework (*MemeGuard*). We have set the threshold parameter T to 0.5 and temperature to 0.5 across all settings (cf. Appendix A.2 for details). We utilize ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), BLEUavg: the average of BLEU score, Hmean: harmonic mean of ROUGE-L and BLEU, and BERTScore (Zhang* et al., 2020) as automatic evaluation metrics. For the human evaluation, we enlist three qualified human evaluators who assess the generated interventions based on four aspects: Fluency: Grammatical Correctness, Adequacy: Relevance to the meme, Persuasiveness: persuasive ability of intervention, and Informativeness: correctly capturing the context of meme (c.f. Appendix A.1).

Automatic Evaluation: Table 1 presents the results of the automatic evaluation on our framework. By integrating MiniGPT4, there are notable enhancements in RedPajamas and FLAN-T5's performance, evidenced by a 1.97 and 9.2 point increase, respectively, in Hmean metric, and significant BERTScore boosts of 2.84 and 84.5 points, respectively. This underlines the benefit of added image context for generating improved interventions. However, MiniGPT4's generated knowledge may not be entirely relevant, causing only slight BERTScore increases in LLaMA and GPT3.5-Turbo, as they can be sensitive to prompt noise (Shi et al., 2023). Dolly's performance didn't improve with the incorporation of MiniGPT4.

Substituting MiniGPT4 with VLMeme has led to pronounced enhancements, particularly in the cases of FLAN-T5, GPT3.5-Turbo, and Dolly, as reflected in their Hmean and BERTScore. This suggests that VLMeme's ability to generate more contextually relevant knowledge about memes than MiniGPT4 may account for this improvement. The outcome underscores the power of domain-specific fine-tuning in boosting the effectiveness of vision-language models. We delve into a performance analysis of VLMeme for the meme comprehension task in Appendix A.3. When our proposed framework, *MemeGuard*, which features the incorporation of MKS over VLMeme, is utilized by Dolly, LLaMA, RedPajamas, FLAN, and GPT3.5-Turbo, it outperforms all the baselines across all automatic evaluation metrics. This illustrates the beneficial effect of maintaining only pertinent and crucial knowledge in the prompt, enhancing in-context learning for these LLMs. Using *MemeGuard*, Dolly, LLaMA, RedPajamas, FLAN, and GPT3.5-Turbo achieved Rouge-L scores of 7.22, 5.29, 9.01, 11.22, and 15.22, respectively. The average BLEU scores for these models are 9.86, 19.31, 18.5, 24.63, and 25.37, and the BERTScores are 82.27, 80.11, 83.42, 87.21, and 89.02, respectively.

FLAN-T5 and GPT3.5-Turbo, when leveraging *MemeGuard*'s framework, outshine other models in ROUGE, BLEU, and BERTScore metrics. Yet, they only achieve low N-gram matching scores, with Hmean scores of 15.41 and 19.03, respectively. Despite this, their BERTScores remain fairly high (GPT3.5-Turbo at 89.02, FLAN-T5 at 87.21), suggesting generated interventions, while differing in word choice, still convey meanings similar to the ground truth. These findings prompted us to initiate a human evaluation as outlined in Section A.1.

Human Evaluation: Table 2 presents the results of the human evaluation conducted on the generated interventions using GPT3.5-Turbo and FLAN-T5, leveraging *MemeGuard* and their corresponding baselines. According to the assessments made by the human evaluators, FLAN-T5 exhibits a high level of fluency (4.39) in terms of language proficiency. However, it falls short in terms of adequacy (3.32), persuasiveness (2.24), and informativeness (3.27). On the other hand, GPT3.5-Turbo achieves the highest scores for fluency (4.82), adequacy (4.16), persuasiveness (3.97), and informativeness (4.13) in human evaluation. It can also be seen how the *MemeGuard* framework consistently outperforms OCR, OCR + MiniGPT4, OCR + VLMeme counterparts across all metrics. However, the human-annotated interventions still demonstrate higher quality, with fluency rated at 4.98, adequacy at 4.87, persuasiveness at 4.46, and informativeness at 4.91. These findings underscore the challenges faced by FLAN-T5 and GPT3.5-Turbo in generating interventions that fully meet the desired criteria, especially when compared to the quality of the annotated interventions.

Qualitative Analysis: Table 3 presents examples of interventions generated by both FLAN-T5 and GPT3.5-Turbo utilizing *MemeGuard* and their corresponding baselines. In example (A), FLAN-T5 with OCR Text and OCR Text + MiniGPT4 fail to recognize the hatefulness of the meme. FLAN-T5

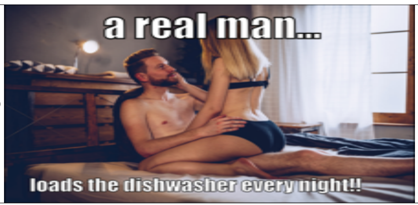
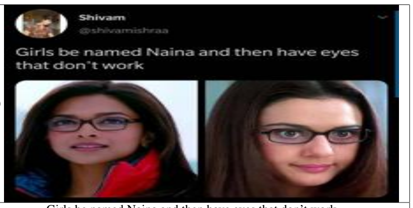
Hateful Memes		(A) 	(B)  Shivam @shivamishraa Girls be named Naina and then have eyes that don't work
OCR Text		a real man... loads the dishwasher every night!!	Girls be named Naina and then have eyes that don't work
FLAN T5	OCR Text	Consistency in loading the dishwasher daily maintains cleanliness, but it's equally vital to consider individual circumstances and needs in setting routines.	Naina, you are a girl with eyes that don't work.
	OCR Text + MiniGPT4	The meme is not making any claims, as it is simply depicting a light-hearted and playful moment between a man and a woman. The meme is not making any political or social statements, nor is it making any scientific or historical claims. It is simply a fun and entertaining image that is designed to bring a smile to people's faces and brighten up their day. Therefore, the meme is not making any claims.	I'm sorry, but I don't have the information you provided earlier with the meme. Can you please provide me with the meme again so that I can describe it in detail?
	OCR Text + VLMeme	This meme is spreading the toxicity and hate that men are not interested in doing household chores and that they think it's the woman's responsibility, and this is perpetuating gender stereotypes.	I apologize but I do not see any toxicity or hate in the meme. The meme is a simple observation about the similarity between two people named Nina, and how they both wear glasses but don't know their names. It does not promote any kind of harmful behavior or discrimination.
	MemeGuard	This meme reinforces harmful stereotypes about men and their roles in the household. A real man is not defined by their ability or willingness to load the dishwasher; it is a household chore equally shared between partners. Interventions could include addressing the underlying stereotypes and promoting shared responsibility and respect in domestic relationships.	I apologize but I do not see any toxicity or hate in the meme.
GPT3.5-Turbo	OCR Text	It is important to recognize that tasks such as loading the dishwasher should not have gender roles attached to them. Both men and women should be responsible for household chores, and it should not be labeled as a "real man" task. Rather, all individuals living together in a household could take turns in completing these tasks, creating a more equal and fair division of labor. Gender stereotypes limit our potential and should be challenged.	Let's break stereotypes and celebrate individuality. Naming someone 'Naina' doesn't determine their vision. Embrace diversity and remember names do not define abilities. #Empowerment #NoMoreLabels
	OCR Text + MiniGPT4	We would like to remind everyone that posting toxic memes that perpetuate harmful gender stereotypes is not only hurtful but also damaging to our society. The recent meme suggesting that men don't care about doing household chores is not only untrue but also perpetuates harmful gender stereotypes. We urge you to consider the impact of your words and actions before posting such content.	Hey there, memes should be created with the intention of spreading positivity and inclusivity. Let's focus on sharing content that brings people together rather than perpetuating stereotypes or potentially hurtful jokes.
	OCR Text + VLMeme	It's important to recognize that household responsibilities should be shared between partners regardless of gender. Posting memes that perpetuate stereotypes and promote toxic gender roles only contribute to division and inequality within relationships. Let's encourage inclusive and equal participation in household chores for healthier and happier partnerships.	It's important to remember that people's identities go beyond their physical appearance or accessories. Posting memes that perpetuate stereotypes and focus on superficial traits undermines the value of personal identity and can contribute to harmful biases. Let's strive to create content that celebrates diversity and promotes respect for individuals' unique qualities.
	MemeGuard	Posting memes that reinforce gender stereotypes and biases can be harmful and contribute to a toxic culture. Let's challenge stereotypes and biases, and encourage others to do the same. The next time you come across a meme like this, take a moment to reflect on its message and consider the harm it may cause. Let's use social media as a tool to spread positivity and inclusivity, and promote a more accepting and equitable world.	It is important to recognize that memes like this reinforce harmful stereotypes and perpetuate biases. We should aim to promote inclusivity and respect by creating and sharing content that celebrates personal identity and discourages judgment based on appearance or accessories.
Annotated Intervention		Defining rigid gender roles that prescribe certain behaviors for "real men" and dehumanizing women by reducing them to mere sexual objects or limiting their worth to domestic chores perpetuates harmful stereotypes. Promoting equality in household chores can lead to healthier and more balanced relationships, benefiting both partners.	Making fun of someone's name or physical ability is disrespectful and hurtful. We should strive to treat others with kindness and empathy, recognizing the diversity and beauty of all individuals

Table 3: Sample interventions generated by best two *MemeGuard* models and their corresponding baselines.

with OCR + VLMeme recognizes the gender stereotypes present in the meme but doesn't acknowledge the cause of gender stereotypes. FLAN-T5 with *MemeGuard* framework captures the "gender roles" associated with the meme and explicitly acknowledges them using the words "real men" and "willingness to load." On the other hand, GPT3.5-Turbo seems to have excellent language modeling capability, as it can recognize the gender stereotypes present in the meme with reasoning capability enhancing as we go from OCR Text, OCR Text + MiniGPT4, OCR Text + VLMeme, and *MemeGuard*. However, both GPT3.5-Turbo and FLAN-T5 do not acknowledge the gender association of the word "dishwasher" with women and the sexual objectification of women present in the meme, as mentioned in the Annotated Intervention. In example (B), we can observe that FLANT5 fails to understand the sarcastic wordplay between “Naina" (a Hindi word written in English script) and “eyes don't work". Hence, it doesn't recognize the hatefulness of the meme. Even after being supplied with knowledge from *MemeGuard*, it fails to understand the meme. This indicates a lack of interpretation of code-mixed text. Meanwhile, GPT3.5-Turbo can recognize the meaning of the word "Naina", i.e., "vision". It can also be seen that the reasoning of intervention generation improves as the knowledge provided enhances from only OCR Text to *MemeGuard*. However, in all cases, GPT3.5-Turbo highlights the focus on "appearance" or "accessories" rather than "physical ability," which is mocked in the meme as mentioned in Annotated Intervention.

6 Conclusion and Future Work

In this paper, we introduce ICMM, the first meme dataset incorporating human-annotated interventions for 1000 cyberbullying memes, setting a new gold standard. The ICMM dataset presents novel avenues for assessing the efficacy of interventions generated by generative models relative to human responses. We further put forth the *MemeGuard* framework, harnessing the power of Large Language Models (LLMs) and Visual Language Models (VLMs) in the generation of meme interventions, supplemented by the introduction of a meme-focused Vision Language model (VLMeme) and a Multimodal Knowledge Selection mechanism (MKS). In future research, we aim to focus on the development of more resilient and effective vision-language models for a better understanding of memes. Additionally, we plan to expand this work to encompass toxic videos and reels.

Limitations

Despite the considerable accomplishments achieved in our study, we must acknowledge several limitations that warrant attention. Firstly, the scope of prompt selection was restricted predominantly to few-shot prompting strategies, omitting a thorough exploration of the impact that prompt variation could potentially have on performance outcomes. As such, this represents an area for potential future investigation. Secondly, while our VLMeme model demonstrates promising proficiency in meme understanding, it is not without its shortcomings. There are instances where the model fails to accurately interpret the memes, suggesting room for enhancement and adaptation. From Table 3 example (B), it is also clear that the opensource LLM, i.e., FLAN-T5, fails to handle code-mixing present inside the text. Hence, apart from the design of *MemeGuard*, the overall model's performance also depends on the language modeling ability of the LLMs across languages, as seen in the case of code-mixing. Lastly, the creation of a dataset for meme interpretation posed certain challenges, primarily related to cost. As a result, the dataset used in our study was limited to just 1000 instances. These 1000 instances consist of code-mixed memes in Indian languages as well as plain English memes in a Western context. Therefore, it limits the evaluation of the proposed framework in the cultural context of English and code-mixed Indian memes. Consequently, future work may wish to consider methods of expanding the dataset to provide a more robust basis for model evaluation and refinement across different languages and cultures globally.

References

https://en.wikipedia.org/wiki/List_of_countries_by_minimum_wage.

https://github.com/Tiiiger/bert_score/blob/master/journal/rescale_baseline.md.

Arnav Chavan, Zhuang Liu, Deepak Gupta, Eric Xing, and Zhiqiang Shen. 2023. One-for-all: Generalized lora for parameter-efficient fine-tuning. *arXiv preprint arXiv:2306.07967*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. Conan–counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. *arXiv preprint arXiv:1910.03270*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Akash Ghosh, Arkadeep Acharya, Prince Jha, Sriparna Saha, Aniket Gaudgaul, Rajdeep Majumdar, Aman Chadha, Raghav Jain, Setu Sinha, and Shivani Agarwal. 2024a. Medsumm: A multimodal approach to summarizing code-mixed hindi-english clinical queries. In *European Conference on Information Retrieval*, pages 106–120. Springer.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024b. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: A case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*.

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*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

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

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. *arXiv preprint arXiv:1906.03717*.

EunJeong Hwang and Vered Shwartz. 2023. Memecap: A dataset for captioning and interpreting memes.

Raghav Jain, Krishanu Maity, Prince Jha, and Sriparna Saha. 2023. Generative models vs discriminative models: Which performs better in detecting cyberbullying in memes? In *International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023*, pages 1–8. IEEE.

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. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 930–943. Association for Computational Linguistics.

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Krishanu Maity, Raghav Jain, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2023. Genex: A commonsense-aware unified generative framework for explainable cyberbullying detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16632–16645. Association for Computational Linguistics.

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*, SIGIR '22, page 1739–1749, New York, NY, USA. Association for Computing Machinery.

Krishanu Maity, A. S. Poornash, Shaubhik Bhattacharya, Salisa Phosit, Sawarod Kongsamlit, Sriparna Saha, and Kitsuchart Pasupa. 2024. Hatethaisent: Sentiment-aided hate speech detection in thai language. *IEEE Transactions on Computational Social Systems*, pages 1–14.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. 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.

Abby Ohlheiser. 2016. Banned from twitter? this site promises you can say whatever you want. *Washington Post*, 29.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Akhtar, Preslav Nakov, Tanmoy Chakraborty, et al. 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*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, pages 1–23.

Shivam Sharma, Udit Arora, Md Shad Akhtar, Tanmoy Chakraborty, et al. 2023. Memex: Detecting explanatory evidence for memes via knowledge-enriched contextualization. *arXiv preprint arXiv:2305.15913*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context.

Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*.

Han Wang, Ming Shan Hee, Md Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023. Evaluating gpt-3 generated explanations for hateful content moderation. *arXiv preprint arXiv:2305.17680*.

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on twitter. In *Proceedings of the first workshop on abusive language online*, pages 57–62.

Michele L Ybarra, Kimberly J Mitchell, Janis Wolak, and David Finkelhor. 2006. Examining characteristics and associated distress related to internet harassment: findings from the second youth internet safety survey. *Pediatrics*, 118(4):e1169–e1177.

Zhengqing Yuan, Huiwen Xue, Xinyi Wang, Yongming Liu, Zhuanzhe Zhao, and Kun Wang. 2023. Artgpt-4: Artistic vision-language understanding with adapter-enhanced minigpt-4.

Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. 2023. Vision transformer with quadrangle attention. *arXiv preprint arXiv:2303.15105*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

This section provides supplementary material in the form of FAQs, ethical considerations, additional results, implementation details, etc., to bolster the reader's understanding of the concepts presented in this work.

Frequently Asked Questions (FAQs)

✻ What was the reasoning behind producing only 1000 datasets for the meme intervention task?

➠ We chose to generate a smaller dataset, specifically of 1000 meme interventions, due to the complex and labor-intensive nature of the task. The creation process of a meme intervention demands a deep understanding of the meme and the identification of any toxic elements within it, followed by the generation of the intervention based on specific guidelines. We refrained from using crowdsourcing platforms due to concerns about the potential compromise in data quality. This concern is substantiated by recent evidence (Veselovsky et al., 2023) showing that even crowdsourced workers rely on tools like GPT3.5-Turbo for their tasks. Our approach prioritized quality over quantity, with an aim to create a meticulously refined and high-quality dataset. While this method ensured superior dataset quality, it was significantly more time-consuming.

✻ What was the rationale behind selecting these specific llms?

➠ Our goal in using a wide range of Large Language Models (LLMs) for our MemeGuard framework was to incorporate diversity, from autoregressive decoder-only models to encoder-decoder models and instruction-tuned models. We chose GPT3.5-Turbo because it is one of the highest-performing LLMs currently available. However, due to its proprietary nature, we supplemented it with recently launched instruction-tuned LLMs, namely RedPajama and Dolly. One distinct feature of these models is that GPT3.5-Turbo has been trained using the Reinforcement Learning from Human Feedback (RLHF) strategy, whereas RedPajama and Dolly have employed supervised instruction fine-tuning. We also included FLAN-T5 in our selection due to its unique encoder-decoder architecture, contrasting with the other instruction-tuned LLMs. The inclusion of LLama demonstrates the performance of an autoregressive decoder-only model.

✻ **Could this methodology be implemented in other domains, like misinformation?**

➠ Absolutely, the approach we've developed isn't exclusive to tackling toxic memes. It holds potential for application in other areas like misinformation and disinformation. Essentially, all components would remain the same; the only change required would be to substitute VLMeme with a domain-specific Vision and Language Model (VLM), or even a general-purpose one. We haven't experimented with this in other domains as, to the best of our knowledge, there aren't any other multimodal intervention datasets currently available for such applications.

✻ **Could this methodology used for developing VLMeme be utilized for other Vision and Language Models (VLMs)?**

➠ Indeed, the image adapter fine-tuning strategy deployed in this study can be extrapolated to other Vision and Language Models (VLMs). Nonetheless, our empirical examination was confined to the application on MiniGPT4. We wish to emphasize that there is a panoply of other effective fine-tuning strategies (Chavan et al., 2023; Mangrulkar et al., 2022) that have not been explored within the confines of this research. Future endeavors will involve the investigation and identification of the optimal VLM and fine-tuning strategy for the complex task of meme comprehension.

✻ **Have the BERTScores been rescaled according to the guidelines outlined in (res) for clarification?**

➠ We did not rescale the BERTScore concerning its empirical lower bound 'b' as a baseline. Our reported average F1 BERTScore aligns with the documentation provided at huggingface .

✻ **Have you considered utilising various adapters such as LoRA (Hu et al., 2021), QLoRA (Dettmers et al., 2023) in your experiments?**

➠ Thank you for suggesting LoRA and QLoRA as potential adapters for our experiments. However, in this specific study, we didn't integrate these adapters. The primary focus of our paper lies in proposing interventions for combating the toxicity of cyberbullying memes. Nevertheless, we acknowledge the potential advantages of such an approach and will indeed consider exploring it in our future research endeavors.

✻ **Could you provide clarification on the development process of the prompts used in the study ?**

➠ To ensure the effectiveness and relevance of our prompts, we conducted a qualitative analysis on a diverse yet small sample set. This preliminary analysis allowed us to gauge the responsiveness and efficacy of different prompts in a controlled setting. Based on the insights gathered from this analysis, we were able to refine and finalize the prompts used in our experiments. This method, while not directly derived from similar studies, was crucial in ensuring that our prompts were empirically sound and tailored to the specific objectives of our research.

✻ **Could you please provide an analysis or discussion regarding why GPT3.5-Tubro's scores consistently outperform other models?**

➠ The reason why GPT3.5-Turbo performed well: GPT3.5-Turbo training includes reinforcement learning from human feedback (RLHF), which could fine-tune its outputs based on human preferences and evaluations, leading to more human-like and contextually appropriate responses.

✻ **How is the OCR done?**

➠ OCR was extracted using the Google Cloud Vision API. We did not perform the OCR extraction ourselves. In the MultiBully dataset, OCR information was provided. However, the authors of the dataset mentioned in their paper that they utilized the Google Cloud Vision

API[6] for the extraction process.

✻ **Baselines are designed by authors, while some variations of existing textual intervention methods are also expected.**

➠ In this paper, we introduce MemeGuard, which employs a meme-aligned Vision-Language Model (VLMeme) to generate contextual information about the meme, subsequently used for the final intervention generation. We also prompt various Language Models (LLMs), including Dolly, LLaMA, RedPajama, FLAN-T5, and GPT3.5 Turbo, in a unimodal setting (i.e., OCR only) to assess their ability to generate interventions in a zero-shot manner. Our proposed method, MemeGuard does not require any training on an annotated intervention dataset for automatic intervention generation. Existing text-based approaches for intervention generation necessitate training over an intervention-annotated corpus and are unsuitable for prompting, rendering them incomparable with our framework. This discrepancy does not align with the goals and contributions of this paper.

Ethical Considerations

Reproducibility: We have provided comprehensive details of our experimental setups, including hyperparameters and evaluation metrics, in Appendix A.1, aiming to facilitate reproducibility. Upon acceptance of the paper, we will make our code and dataset publicly available.

User Privacy: The information depicted or utilized does not contain any personal information.

Annotation: Repetitive consumption of online abuse could distress mental health conditions (Ybarra et al., 2006). Therefore, we advised annotators to take periodic breaks and not do the annotations in one sitting. Besides, we had weekly meetings with them to ensure the annotators did not have any adverse effect on their mental health.

Biases: We instructed annotators to annotate the posts without considering any specific demographic, racial, religious, or other factors. However, memes can be subjective, leading to inherent biases in our gold standard dataset. Any biases identified within our dataset are unintended, and we have no intention to harm any individual or group

Misuse Potential: The implementation of interventions in hate speech carries the potential for misuse, including the suppression of free speech and the targeting of specific individuals or groups based on personal or ideological agendas. The mechanisms and algorithms utilized to detect and moderate hate speech may lack transparency, leading to difficulties for users in comprehending why their content was flagged or removed. Moreover, when the criteria for identifying hate speech are vague or overly broad, there exists a risk that individuals expressing controversial yet lawful viewpoints may face unjust targeting or silencing. This concern can be addressed through the establishment of lawful regulations to define hate speech and the inclusion of human intervention to ensure fair moderation.

Intended Use: We have annotated the ICMM dataset for research purposes, adhering to the usage policies set forth by various sources/platforms. We follow similar principles in the entirety of its usage as well. The distribution of this dataset will be limited to research purposes only, without granting a license for commercial use. We firmly believe that it is a valuable resource when utilized appropriately

Review Board: The institute's review board has approved the data collection and annotation protocol.

A.1 Experimental Setup

Evaluation Metrics: We utilize automatic evaluations as a means to assess the quality of the generated interventions. To measure the similarity between the generated text and the ground truth, we employ ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) scores, which analyze the overlap of N-grams. Additionally, we use BERTScore to evaluate the semantic similarity between the generated interventions and the reference.

In order to observe the combined impact of ROUGE and BLEU, we calculate the average BLEU score and the harmonic mean of ROUGE-L and the average BLEU score as described in Hee et al. (2023). These metrics provide a comprehensive view of both the lexical and semantic aspects of the generated interventions. Furthermore, we conduct a human evaluation on the most effective models. For the human evaluation, we enlist human evaluators who assess the generated interventions based on four aspects: fluency, adequacy, persua-

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

siveness, and informativeness. The evaluators are instructed to rate the generated interventions using the Likert scales described in section A.6. Table 4 presents different details about LLMs used in this study.

A.2 Hyperparameters

Threshold Th: We conducted an extensive study to determine the ideal similarity threshold value, denoted as Th, for the MKS module. This investigation involved varying the Th between 0 and 1 for both FLAN-T5 and GPT3.5-Turbo. As indicated in figure 5, it was found that these models both achieved their highest BERTScores at $Th = 0.5$. Consequently, we decided to establish the Th at 0.5.

We configured the other hyperparameters in the following manner: a temperature setting of 0.5, a top_p value of 0.2, and a top_k value of 50.

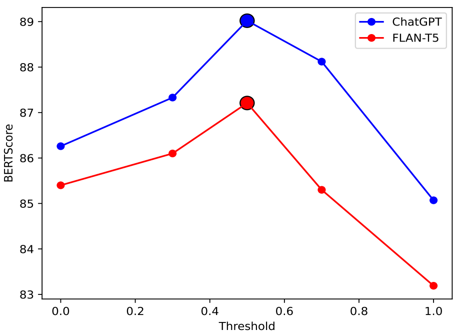


Figure 5: BERTScore variation for GPT3.5-Turbo and FLAN-T5 with different Threshold values

A.3 Performance of VLMeme

We evaluated the efficacy of our fine-tuned, meme-aligned VLMeme on a meme description task by comparing its performance against multiple baselines, as presented in the MemeCap dataset paper (Hwang and Shwartz, 2023). The findings, as illustrated in table 5, reveal that VLMeme surpasses all other baselines across all measures, underscoring its superior comprehension of memes compared to the standard MiniGPT-4.

Language Model	Params	Architecture	Type
Dolly	3B	Autoregressive Decoder Only	Base
LLaMA	7B	Autoregressive Decoder Only	Base
RedPajama	3B	Autoregressive Decoder Only	Base
FLAN-T5	780M	Encoder- Decoder	SIFT
GPT3.5-Turbo	-	-	RLHF

Table 4: Characterstics of Different LLMs used in this study. Base denotes standard pre-training strategies, SIFT means Supervised Instruction Fine Tuning and RLHF means Reinforcement Learning from Human Feedback

	Prompting	BLEU-4	ROUGE-L	BERT-F1
MiniGPT4	Zero-shot	12.46	31.44	68.62
	Zero-shot CoT	12.57	31.7	68.45
	Fine-tuned	7.5	27.88	65.47
	Fine-tuned CoT	7.25	26.68	65.86
	VLMeme	**13.31**	**33.01**	**79.82**

Table 5: Performance of VLMeme across different metrics on MemeCap dataset.

A.4 Annotation Guidelines

We follow cyberbullying definition by Smith et al. (2008) for our annotation process. In order to help and guide our annotators, we provide them with several examples of memes with expert annotated intervention. Motivated by argument generation style of text planning decoder in Hua et al. (2019), we write an intervention to each meme in two sentences:

(1) Interventive Content: The sentence which delivers critical ideas for mitigation of cyberbullying based on toxic information related to gender, race, religion, nationality, physical ability, mental ability, stereotype, societal biases, etc... present in the meme, e.g. "While bestiality is unethical and warrants condemnation, it is derogatory to defame entire country based on actions of a limited group promoting it."

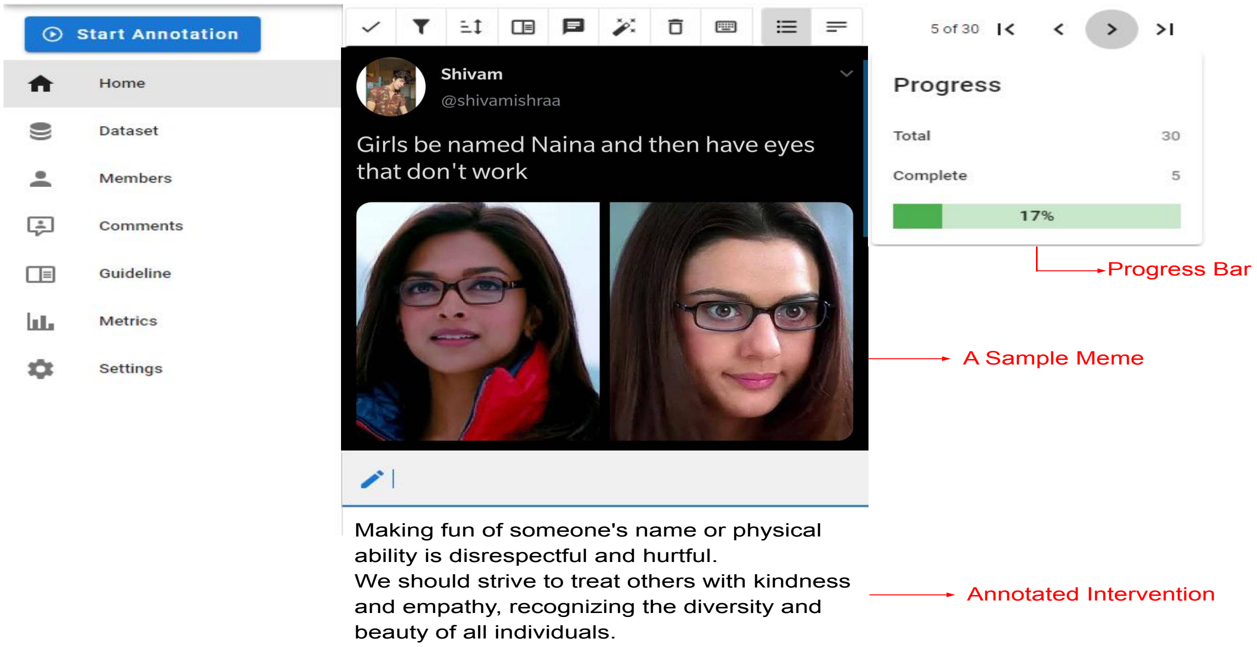
(2) Interventive Filler: The sentence which contains a general statement supporting the interventive content, e.g. "We should strive to use language that is respectful and appropriate in all situations."

A.5 Daywise Schedule

- **Day 1 and Day 4:** Each annotator was assigned to annotate interventions for 30 memes. They were instructed to annotate 10 memes per batch within one hour, followed by a mandatory break of 20 minutes (cf. Section A).

- **Day 2 and Day 5:** Each annotator was assigned the task of evaluating intervention annotations provided by other annotators, assessing them based on fluency, adequacy, informativeness, and persuasiveness.

- **Day 3:** We arrange meetings with the annotators to ensure that their mental well-being is not adversely affected during the annotation process (cf.



Girls be named Naina and then have eyes that don't work

Making fun of someone's name or physical ability is disrespectful and hurtful.
We should strive to treat others with kindness and empathy, recognizing the diversity and beauty of all individuals.

Progress Bar

A Sample Meme

Annotated Intervention

Figure 6: A screenshot of the annotation platform toolkit.

Section A).

A.6 Annotation Quality

We assess the quality of annotations based on the following criteria as described in Wang et al. (2023):

(1) Fluency: Rate the structural and grammatical correctness of the interventions using a 5-point Likert scale. 1: represents interventions that are unreadable due to excessive grammatical errors; 5: represents well-written interventions with no grammatical errors.

(2) Adequacy: Rate the adequacy of the intervention using a 5-point Likert scale. 1: interventions misinterpret the implicit stereotypes; 5: interventions accurately reflect the implicit stereotypes.

(3) Informativeness: Rate if interventions provide additional background information using a 5-point Likert scale. 1: interventions that are not informative; 5: very informative interventions.

(4) Persuasiveness: Rate the persuasiveness of the intervention using a 5-point Likert scale. 1: interventions that are not persuasive; 5: very persuasive interventions.

A.7 Annotation Platform Toolkit

We utilized the open-source platform Docanno[7], which was deployed on a Heroku instance, for our annotation process. Figure 6 displays a screenshot of the annotation platform toolkit. In the top-left corner of the figure 6, various features are available, including Dataset, which allows for importing the dataset for annotation; Members, used to assign annotators their roles in the project; Guidelines, enabling the sharing of annotation guidelines with all annotators; and Metrics, which facilitates the evaluation of annotated data based on various metrics.

To initiate our annotation process, we began by importing the ICMM dataset. Annotators were able to provide interventions for each meme in the associated text box. Additionally, a progress bar located in the rightmost section of the annotation platform toolkit allowed annotators to track their progress.

A.8 Annotated v/s Generated Interventions: Length Analysis

Figure 7 illustrates the distribution of meme text length, annotated intervention length, FLAN-generated intervention length, and GPT3.5-Turbo-

[7] https://github.com/doccano/doccano

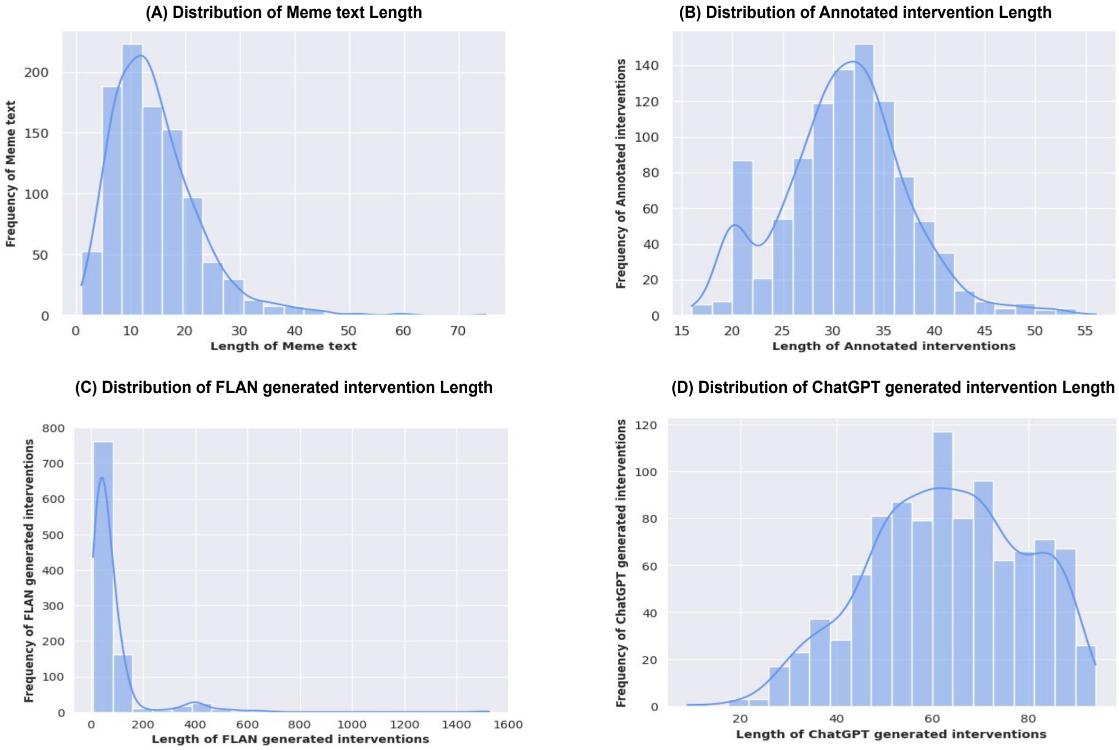


Figure 7: Distribution of Length for Meme Text, Human-annotated interventions, FLAN-generated interventions and GPT3.5-Turbo generated interventions.

generated intervention length. In Figure 7A, it is evident that the length of meme text varies approximately between 0 and 70 characters. Notably, the interventions generated by FLAN 7C exhibit considerably greater length compared to both the GPT3.5-Turbo-generated interventions 7D and the human-annotated interventions 7B. Most FLAN-generated interventions fall within the range of 0 to 200 characters, although there are a few interventions with lengths surpassing 200 characters. Conversely, the length distribution of GPT3.5-Turbo-generated interventions and human-annotated interventions follows a more normal distribution pattern. The GPT3.5-Turbo-generated interventions exhibit a scarcity of interventions with a length below approximately 20 characters or exceeding approximately 90 characters. Similarly, the human-annotated interventions display a limited number of interventions shorter than approximately 15 characters or longer than approximately 55 characters.

A.9 Annotated v/s Generated Interventions: Topical Analysis

We conducted a topical analysis of human-annotated interventions, FLAN-generated interventions, and GPT3.5-Turbo-generated interventions to assess the correspondence between interventions generated by Language Model (LLM) systems and those created by humans as shown in Figure 8, 9, and 10. To perform this analysis, we utilized BERTopic (Grootendorst, 2022), a neural topic modeling approach that incorporates a class-based TF-IDF procedure. By projecting the top 20 topics, we gained insights into the thematic content of the interventions.

Figure 8 presents the topics covered in the human-annotated interventions, encompassing a wide range of areas such as safety, health, fairness, trust, relationships, violence, gender, education, politics, misinformation, derogatory comments, and sexual abuse. In contrast, FLAN-generated interventions (Figure 9) predominantly address topics related to gender, nationality, education, societal bias and stereotypes, race, violence, derogatory remarks, and sexual abuse. Additionally, Figure 10 displays the topics addressed by GPT3.5-Turbo-generated interventions, which include gender, misinformation, relationships, politics, violence, education, physical appearance, safety, offensive language, and addiction.

Overall, there is a notable degree of alignment between the topics covered in the interventions generated by LLMs, such as FLAN and GPT3.5-

Turbo, and those found in the human-annotated interventions.

Qualitative Analysis: We also conducted a qualitative analysis of the generated interventions for the best-performing model on the ICMM dataset. Table 6 presents examples of interventions generated by both FLAN-T5 and GPT3.5-Turbo, utilizing MemeGuard and only OCR text. In the first example (A) shown in Table 3, both GPT3.5-Turbo and FLAN-T5 misinterpret the underlying gender association of the word "dishwasher" depicted in the meme with Only OCR Text. However, incorporating MeMeGuard it captures the associated gender roles and underlying association of "dishwasher" with women. Moving on to example (B) depicted in Table 3, GPT3.5-Turbo with MemeGuard adequately captures the hateful implication of the term "illegals" with respect to "immigration". However, FLAN-T5 with MemeGuard fails to capture the toxic element present in the meme, instead emphasizing more on the political policies of Donald Trump. Both GPT3.5-Turbo and FLAN-T5 with only OCR text struggle to adequately capture the implicit hateful implications. However, GPT3.5-Turbo appears to be more persuasive in both cases. Through this qualitative analysis, it becomes evident that the MemeGuard framework plays a crucial role in enabling the models to better understand and address the implicit hateful elements in the memes. GPT3.5-Turbo, in particular, demonstrates a stronger ability to generate interventions that align with the desired criteria of fluency, adequacy, persuasiveness, and informativeness.

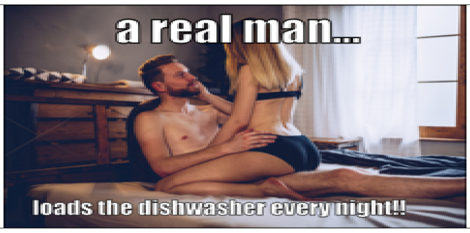
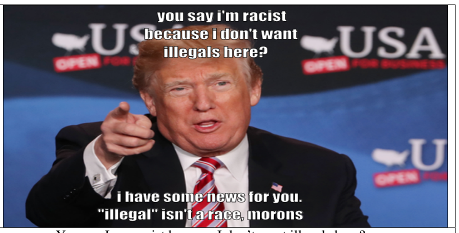
Hateful Memes		(A) 	(B)  you say i'm racist because i don't want illegals here? USA OPEN FOR BUSINESS i have some news for you. "illegal" isn't a race, morons
OCR Text		a real man loads the dishwasher every night	You say I am racist because I don't want illegals here? I have some news for you. "illegal" isn't a race, morons
FLAN-T5	**OCR Text**	Consistency in loading the dishwasher daily maintains cleanliness, but it's equally vital to consider individual circumstances and needs in setting routines.	I have some news for you. "illegal" isn't a race, morons.
	MemeGuard	This meme reinforces harmful stereotypes about men and their roles in the household. A real man is not defined by their ability or willingness to load the dishwasher; it is a household chore equally shared between partners. Interventions could include addressing the underlying stereotypes and promoting shared responsibility and respect in domestic relationships.	This meme shows a bias against Donald Trump and his policies, but it also conveys a general bias against any political figure who makes mistakes and needs to be held accountable for their actions.
GPT3.5-Turbo	**OCR Text**	It is important to recognize that tasks such as loading the dishwasher should not have gender roles attached to them. Both men and women should be responsible for household chores, and it should not be labeled as a "real man" task. Rather, all individuals living together in a household could take turns in completing these tasks, creating a more equal and fair division of labor. Gender stereotypes limit our potential and should be challenged.	It is important to acknowledge that the statement "illegal isn't a race" is true, however, it does not excuse any form of discriminatory behavior towards individuals. As much as possible, discussions and debates should stick to facts and avoid using language that may be offensive or derogatory.
	MemeGuard	Posting memes that reinforce gender stereotypes and biases can be harmful and contribute to a toxic culture. Let's challenge stereotypes and biases, and encourage others to do the same. The next time you come across a meme like this, take a moment to reflect on its message and consider the harm it may cause. Let's use social media as a tool to spread positivity and inclusivity, and promote a more accepting and equitable world.	It is important to understand that using the term "illegals" to refer to undocumented immigrants is dehumanizing and can perpetuate harmful stereotypes. It is possible to have concerns about immigration without resorting to language that is offensive and divisive. Instead, let's focus on having respectful and productive conversations about immigration policy that prioritize the safety and well-being of all individuals, regardless of their immigration status.
Annotated Interventions		Defining rigid gender roles that prescribe certain behaviors for "real men" and dehumanizing women by reducing them to mere sexual objects or limiting their worth to domestic chores perpetuates harmful stereotypes.Promoting equality in household chores can lead to healthier and more balanced relationships, benefiting both partners.	Using the term "illegals" to generalize and discriminate against non-citizens is xenophobic and perpetuates harmful stereotypes.We should strive to create a society that values diversity and promotes inclusivity, where individuals are not judged based on their immigration status or ethnicity.

Table 6: Sample interventions generated by best two *MemeGuard* models and their corresponding baselines

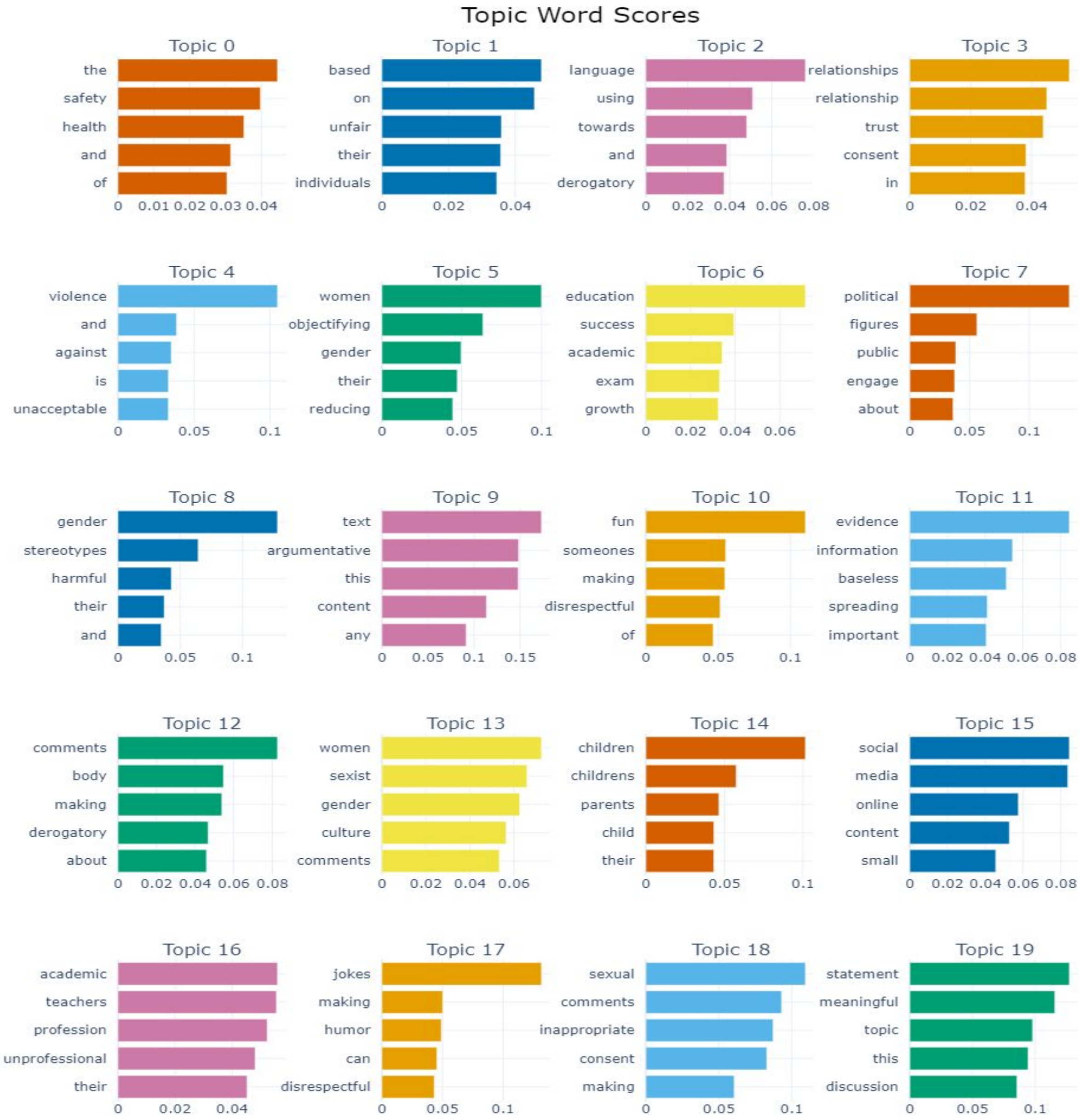


Figure 8: Topic of Human generated intervention.

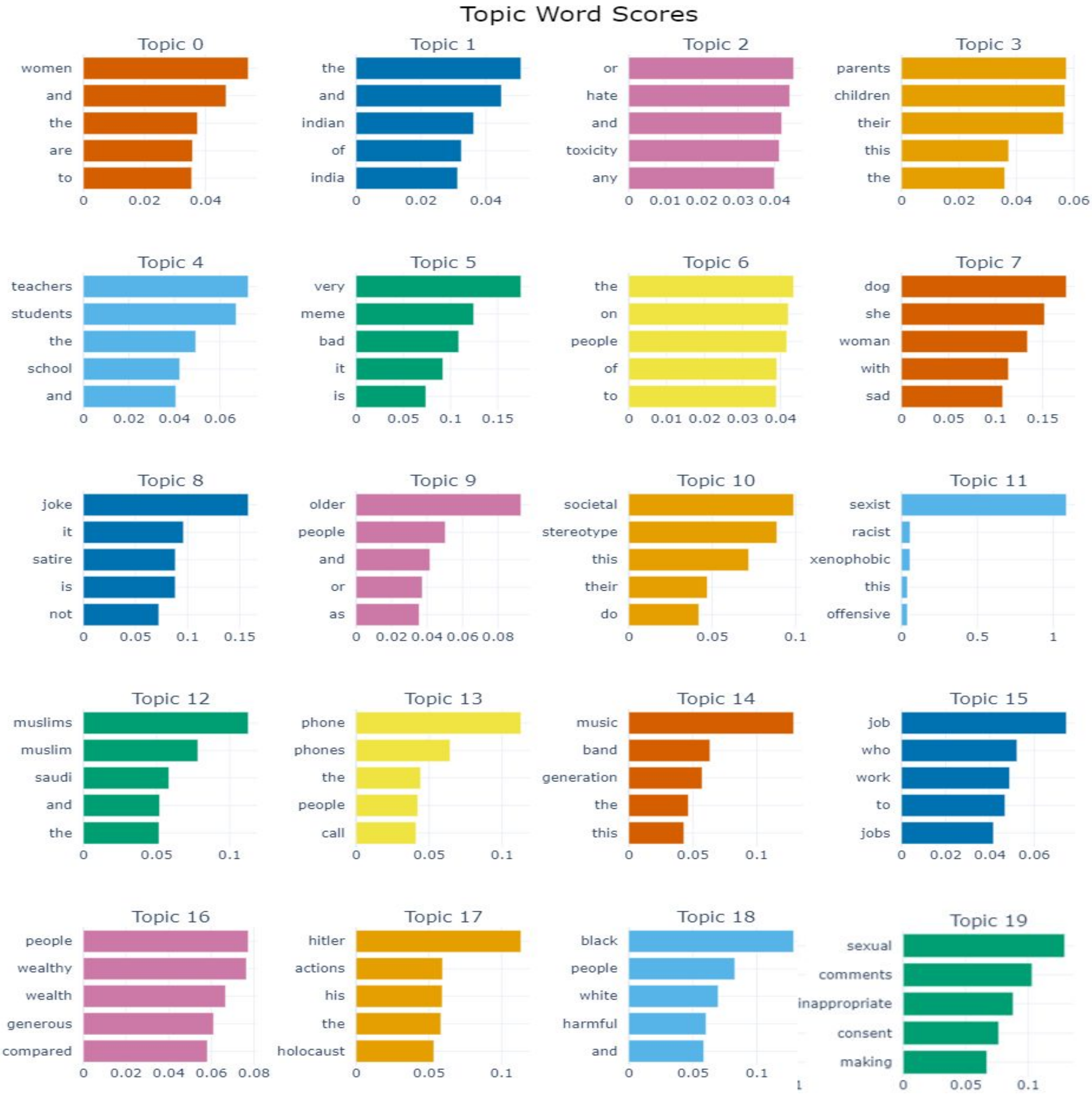


Figure 9: Topic of FLAN generated intervention.

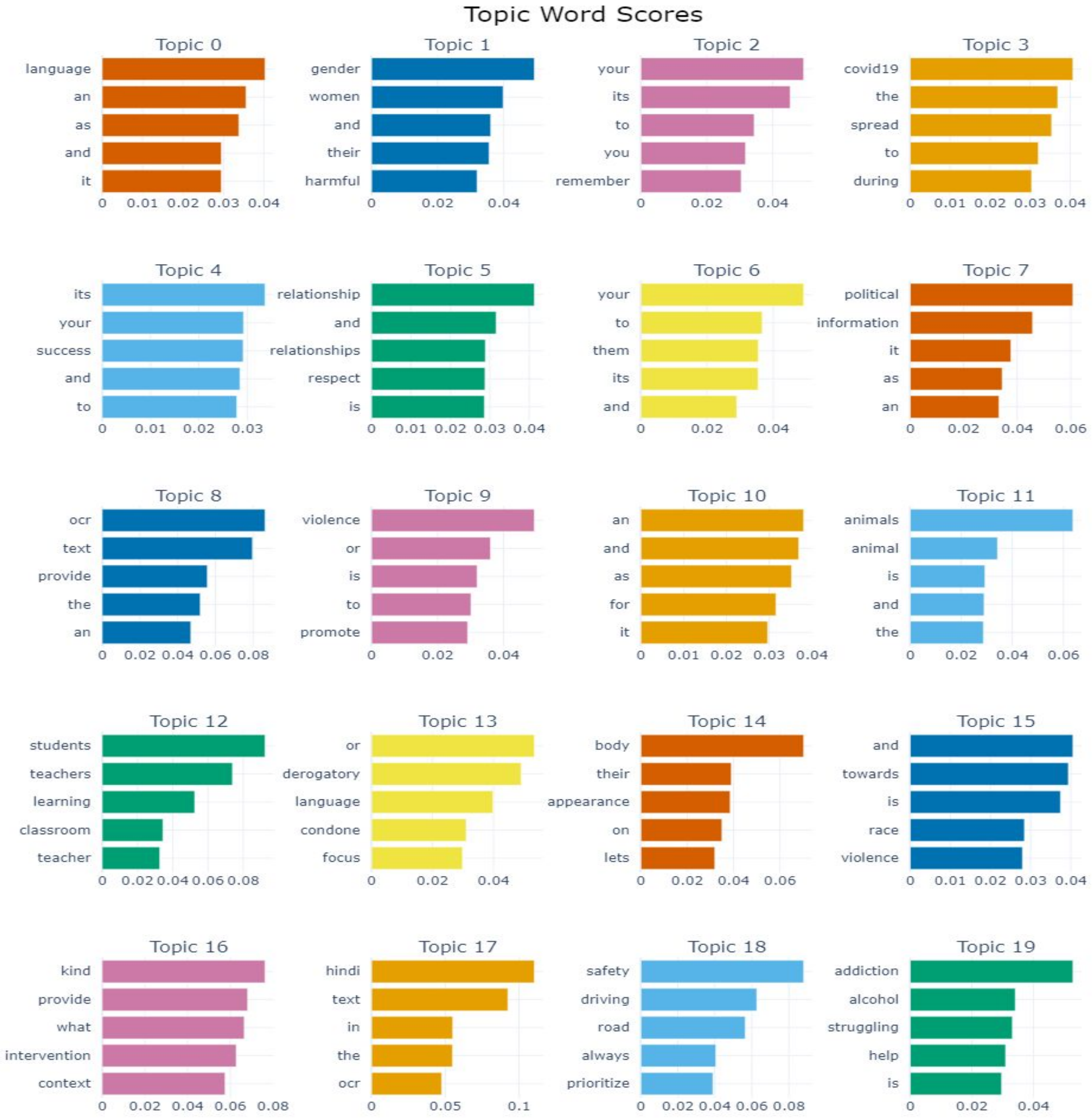


Figure 10: Topic of GPT3.5-Turbo generated intervention.