I know what you MEME! Understanding and Detecting Harmful Memes with Multimodal Large Language Models

Authors: Yong Zhuang, Keyan Guo, Juan Wang, Yiheng Jing, Xiaoyang Xu, Wenzhe Yi, Mengda Yang, Bo Zhao, Hongxin Hu

Disclaimer: This presentation contains harmful content, which has the potential to be offensive and may disturb readers.

Presented by Internet Society

NDSS SYMPOSIUM

NDSS SYMPOSIUM

What is a Meme?

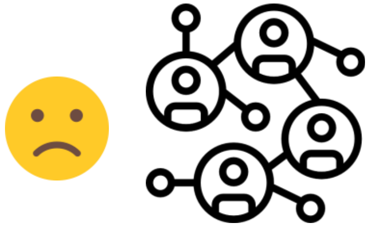
- A meme is a viral social media format that combines images and text to convey ideas, humor, or culture.

NDSS SYMPOSIUM

The Dark Side of Memes - Harmful Meme

Malicious Users

Social Platform



Online Users

NDSS SYMPOSIUM

Challenges in Harmful Meme Detection

- **Multimodal semantic fusion**

  The interplay between text and images in memes can convey both subtle and overt harmful content.

- **Meme composition**

  The arrangement of visual elements influences perception, potentially masking harmful intent.

- **Meme propaganda techniques**

  The use of strategic rhetorical and psychological tactics can influence opinions or behaviors, potentially obscuring harmful content.

Background

Motivation

Method

Experiment

Conclusion

Challenge 1 - Multimodal semantic fusion

➢ Dataset

HatReD[5], contains human semantic annotations of harmful memes.

➢ Measurement

Evaluation: BERTScore

$$R_{\mathrm{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^{\top} \hat{x}_j$$

$$P_{\mathrm{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^{\top} \hat{x}_j$$

$$F_{\mathrm{BERT}} = 2 \cdot \frac{P_{\mathrm{BERT}} \cdot R_{\mathrm{BERT}}}{P_{\mathrm{BERT}} + R_{\mathrm{BERT}}}$$

x - human interpretation of the meme

$\hat{x}$ - model's interpretation of the meme

x_i, $\hat{x}_j$ - Token embeddings from BERT

$x_i^{\top}\hat{x}_j$ - similarity between token embeddings

[5] M. S. Hee, W.-H. Chong, and R. K.-W. Lee, "Decoding the underlying meaning of multimodal hateful memes," in Proceedings of the ThirtySecond International Joint Conference on Artificial Intelligence, 2023, pp. 5995–6003.

Background

Motivation

Method

Experiment

Conclusion

Challenge 1 - Multimodal semantic fusion

➢ Dataset

HatReD[5], contains human semantic annotations of harmful memes.

➢ Measurement

Evaluation: BERTScore

$$R_{\mathrm{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j$$

$$P_{\mathrm{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^\top \hat{x}_j$$

$$F_{\mathrm{BERT}} = 2 \cdot \frac{P_{\mathrm{BERT}} \cdot R_{\mathrm{BERT}}}{P_{\mathrm{BERT}} + R_{\mathrm{BERT}}}$$

Model	BERTScore		
	P_{BERT}	R_{BERT}	F_{BERT}
VisualBERT	0.5	0.45	0.47
VL-T5	0.47	0.41	0.45
LLaVA	0.77	0.80	0.79
GPT-4	**0.84**	**0.83**	**0.83**

Remark 1 - The challenge of multimodal semantic fusion for traditional models, and the ability of MLLM in meme understanding

x - human interpretation of the meme

$\hat{x}$ - model's interpretation of the meme

x_i, $\hat{x}_j$ - Token embeddings from BERT

$x_i^\top \hat{x}_j$ - similarity between token embeddings

[5] M. S. Hee, W.-H. Chong, and R. K.-W. Lee, "Decoding the underlying meaning of multimodal hateful memes," in Proceedings of the ThirtySecond International Joint Conference on Artificial Intelligence, 2023, pp. 5995–6003.

Challenge 2 - Meme Composition

From the perspective of visual arts, meme composition is the arrangement of visual elements

- **Number of Panels**
- **Type of the Images**
- **Scale**
- **Movement**

Challenge 2 - Meme Composition

- **Number of Panels**

Single

Stitching

Challenge 2 - Meme Composition

➢ **Type of the Images**

Photo

Screenshot

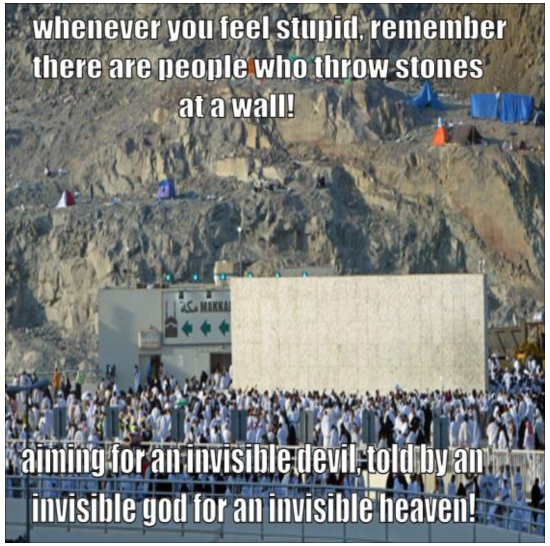
Illustration

Challenge 2 - Meme Composition

➢ **Scale**

Close up

Medium Shot



Long Shot

Challenge 2 - Meme Composition

- **Movement**

Physical

Emotional

Casual

Background
Motivation
Method
Experiment
Conclusion

Challenge 2 - Meme Composition

Category	Subcategory	Proportion of Meme	True Positive Rate (TPR)			
			ExplainHM	LLaVa	GPT-4	Avg.
Number of Panels	Single	67%	67.03%	18.18%	61.54%	52.35%
	Stitching	33%	42.42%	16.48%	33.3%	35.18%
Type of the Images	Illustration	17%	71.43%	14.29%	66.67%	57.15%
	Photo	50%	63.04%	17.39%	58.7%	51.09%
	Screenshot	33%	73.91%	17.39%	65.22%	54.35%
Scale	Close-up shot	20%	75%	42.86%	85.71%	68.75%
	Medium shot	76%	63.64%	12.99%	59.74%	50%
	Long shot	4%	80%	0%	0%	40%
Movement	Physical movement	56%	61.36%	18.18%	63.64%	52.27%
	Emotional movement	39%	67.74%	16.13%	61.29%	52.42%
	Causal movement	5%	75%	0%	25%	37.5%

Challenge 2 - Meme Composition

Category	Subcategory	Proportion of Meme	True Positive Rate (TPR)			
			ExplainHM	LLaVa	GPT-4	Avg.
Number of Panels	Single	67%	67.03%	18.18%	61.54%	52.35%
	Stitching	33%	42.42%	16.48%	33.3%	35.18%
Type of the Images	Illustration	17%	71.43%	14.29%	66.67%	57.15%
	Photo	50%	63.04%	17.39%	58.7%	51.09%
	Screenshot	33%	73.91%	17.39%	65.22%	54.35%
Scale	Close-up shot	20%	75%	42.86%	85.71%	68.75%
	Medium shot	76%	63.64%	12.99%	59.74%	50%
	Long shot	4%	80%	0%	0%	40%
Movement	Physical movement	56%	61.36%	18.18%	63.64%	52.27%
	Emotional movement	39%	67.74%	16.13%	61.29%	52.42%
	Causal movement	5%	75%	0%	25%	37.5%

Remark 2 - Meme composition challenges harmful meme detection particularly with stitched images, which complicate understanding visually.

Challenge 3 - Propaganda Techniques

The use of strategic rhetorical and psychological tactics can influence opinions or behaviors, potentially obscuring harmful content.

- Name calling or labeling
- Appeal to fear/ prejudices
- Whataboutism
- Misrepresentation of someone's position
- Flag-waving
- Causal oversimplification
- Black-and-white fallacy or dictatorship
- Reductio ad hitlerum
- Smears
- Loaded language
- Doubt
- Exaggeration/ Minimisation
- Slogans
- Appeal to authority
- Thought-terminating cliche
- Repetition
- Obfuscation, Intentional vagueness, Confusion
- Presenting irrelevant data
- Bandwagon
- Glittering generalities
- Appeal to strong emotions
- Transfer

[6] D. Dimitrov, B. Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G. Da San Martino et al., "Detecting propaganda techniques in memes," in ACL-IJCNLP 2021-59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference. Association for Computational Linguistics (ACL), 2021, pp. 6603–6617.

NDSS SYMPOSIUM

- Background
- Motivation
- Method
- Experiment
- Conclusion

Challenge 3 - Propaganda Techniques

➢ **Appeal to strong emotions:**
Uses emotionally charged imagery to evoke fear/anger, linking modern Democrats with extreme historical groups

➢ **Name-calling:**
Labels Democrats as "extremists" or "racists" to provoke negative associations

➢ **Smears:**
Damages the reputation of Democrats by associating them with negative stereotypes

➢ **Transfer:**
Associates negative qualities of historical groups with modern Democrats

Challenge 3 - Propaganda Techniques

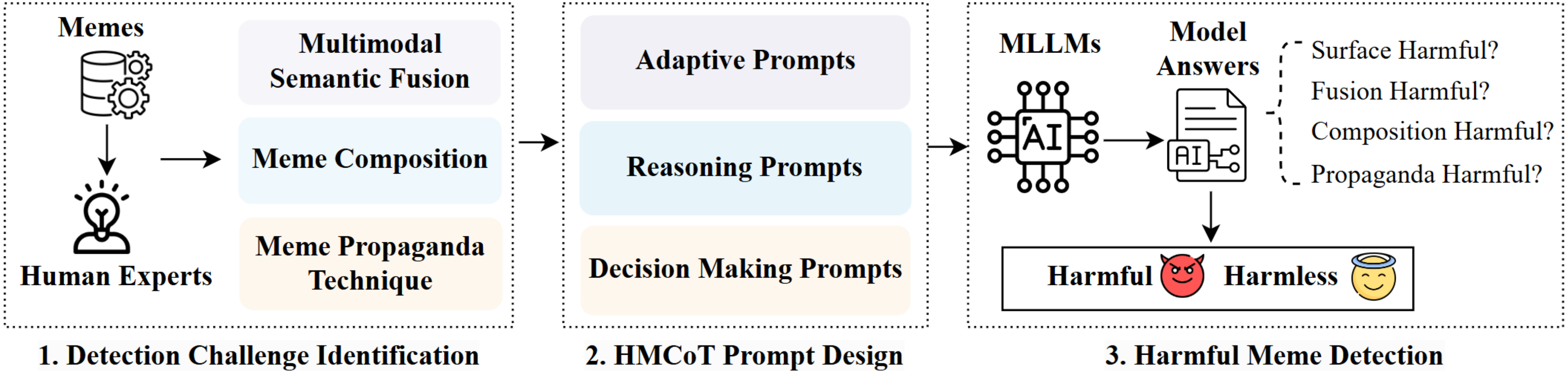
NDSS SYMPOSIUM

- Background
- **Motivation**
- Method
- Experiment
- Conclusion

Catetory	Proportion of Meme	True Positive Rate (TPR)			
		ExplainHM	LLaVa	GPT-4	Avg.
w/o propaganda techniques	57.3%	75%	17.31%	60%	50.77%
w/ propaganda techniques	42.7%	53.85%	15%	48.08%	38.98%

Remark 3 - Meme propaganda technique poses challenges for detecting harmful content, as it makes the expression more subtle and less detectable.

Overview of HMGuard



NDSS SYMPOSIUM

Overview of HMCoT

Adaptive Prompt

Adaption: This is a meme, using text and images for humor or satire, shaped by cultural and contexts. You are a content moderation specialist. Your task is to pinpoint any instances of hate speech, explicit violence, discrimination, or any other type of content that may be considered harmful.

Reasoning Prompt

Q1: What are the words and images contained in the meme, and do the semantics of the text and image directly convey harmful content?

Q2: What are the explicit or implicit relationships between text and image, and does the relationship between the text and the image potentially reveal harmful content?

Q3a: Is the meme a stitching image?

Yes

Q3b: Consider the relationship of images, and understand whether the stitching images are trying to express explicit or implicit harmful content.

Q4: Are the following propaganda techniques used to explicitly or implicitly express harmful content?

Q5: Does the meme intend to have any targeted derogatory, humiliating, insulting, satirical, or disparaging meaning?

Decision Making Prompt

Final Decision: Combining the analysis from the previous questions, please make the final decision on whether this meme is harmful or harmless. You need to make sure that your answers are consistent with the questions above.

Evaluation

➢ Dataset

Dataset	# Memes	# Harmful	# Harmless
HarMeme	289	110	179
FHM	711	422	289
Total	1000	532	468

➢ Baselines

(1) MOMENTA [24]: Multimodal harmful meme detection system, released the Harmeme dataset.

(2) HateDetectron [48]: Wining the Meta's Hateful Meme Challenge with FHM dataset.

(3) MR.HARM [14]: LLM-based harmful meme detection system.

(4) ExplainHM [29]: Uses LLMs' argumentative abilities for diverse explanations, achieving state-of-the-art detection.

(5) GPT-4 [16]: Advanced MLLM with strong reasoning capabilities.

NDSS SYMPOSIUM

Evaluation

➢ **Compare to baselines**

Detector	FHM				HarMeme			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
MOMENTA	0.61	0.66	0.51	0.57	0.77	0.69	0.45	0.55
HateDetectron	0.69	0.54	0.73	0.58	0.8	0.77	0.62	0.68
MR.HARM	0.58	0.34	0.65	0.45	0.8	0.56	0.82	0.66
ExplainHM	0.48	0.35	0.55	0.48	0.73	0.25	0.62	0.71
GPT-4	0.61	0.55	0.64	0.6	0.74	0.72	0.5	0.69
HMGuard	**0.86**	**0.88**	**0.83**	**0.85**	**0.92**	**0.83**	**0.98**	**0.91**

Note: Underline represents the best results in baselines; **Bolding** represents the best results among all approaches.

Effectiveness of addressing the challenges

Category of Harmful Meme		HMGuard	Improvement
Number of Panels	Single	97.44%	30.41%
	Stitching images	96.88%	54.46%
Type of the Images	Illustration	99.99%	28.56%
	Photo	97.44%	34.40%
	Screenshot	95%	21.09%
Scale	Close-up shot	99.99%	14.28%
	Medium shot	96.92%	33.28%
	Long shot	99.99%	19.99%
Movement	Physical movement	94.59%	30.95%
	Emotional movement	99.99%	32.25%
	Causal movement	99.99%	24.99%
w/o propaganda techniques		99.99%	24.99%
w/ propaganda techniques		97.78%	43.93%

Detecting “in-the-wild” Harmful Memes

Detector	Accuracy	Percision	Recall	F1-score
HateDetectron	0.7	0.53	0.52	0.51
MR.HARM	0.73	0.57	0.52	0.5
HMGUARD	**0.88**	**0.83**	**0.89**	**0.86**

Pinterest

Conclusion

➢ Contributions

- New understanding of harmful memes from novel perspectives
- New framework for harmful meme detection.
- Extensive evaluation of HMGUARD.

➢ Future Work

- Multilingual harmful meme detection
- Promising approaches such as RAG and AI Agent

Thanks for your attention!

Q&A

Wuhan University - Trusted Computing & Security Lab

Yong Zhuang yong.zhuang@whu.edu.cn

University at Buffalo

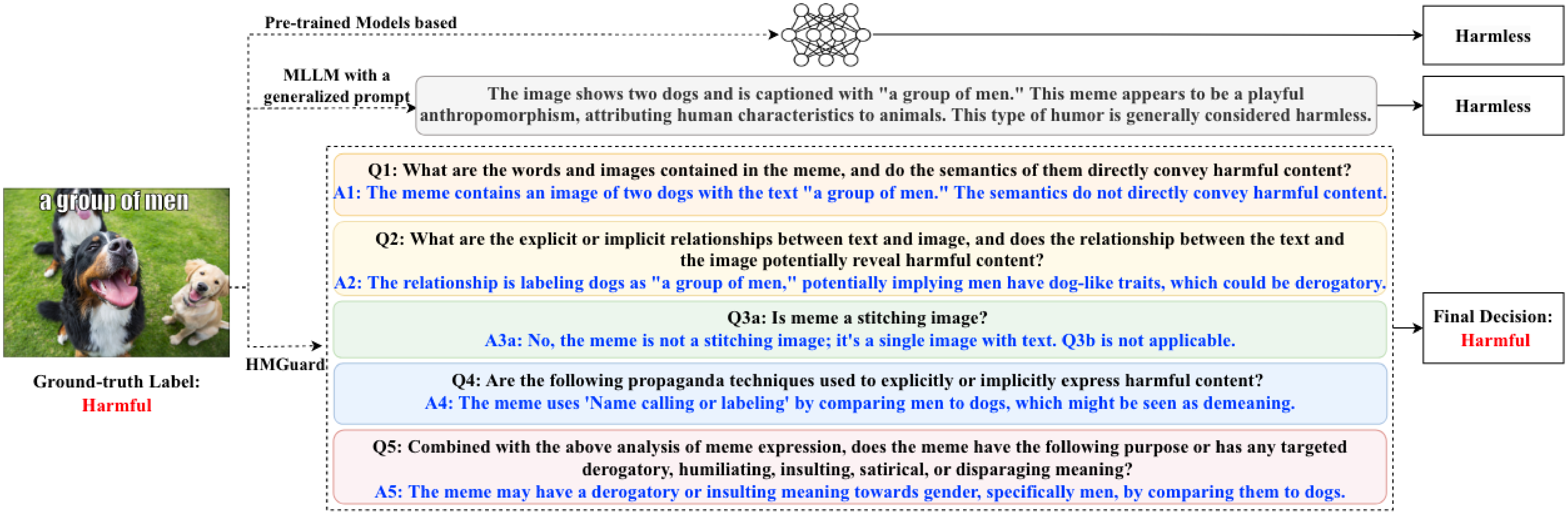
Keyan Guo keyanguo@buffalo.edu

Presented by

Internet Society

NDSS SYMPOSIUM

Example



a group of men
Ground-truth Label:
Harmful
Pre-trained Models based
Harmless
MLLM with a generalized prompt
The image shows two dogs and is captioned with "a group of men." This meme appears to be a playful anthropomorphism, attributing human characteristics to animals. This type of humor is generally considered harmless.
Harmless
HMGuard
Q1: What are the words and images contained in the meme, and do the semantics of them directly convey harmful content?
A1: The meme contains an image of two dogs with the text "a group of men." The semantics do not directly convey harmful content.
Q2: What are the explicit or implicit relationships between text and image, and does the relationship between the text and the image potentially reveal harmful content?
A2: The relationship is labeling dogs as "a group of men," potentially implying men have dog-like traits, which could be derogatory.
Q3a: Is meme a stitching image?
A3a: No, the meme is not a stitching image; it's a single image with text. Q3b is not applicable.
Q4: Are the following propaganda techniques used to explicitly or implicitly express harmful content?
A4: The meme uses 'Name calling or labeling' by comparing men to dogs, which might be seen as demeaning.
Q5: Combined with the above analysis of meme expression, does the meme have the following purpose or has any targeted derogatory, humiliating, insulting, satirical, or disparaging meaning?
A5: The meme may have a derogatory or insulting meaning towards gender, specifically men, by comparing them to dogs.
Final Decision:
Harmful