Supplementary Material

A Data Annotation Details

In general, human reasoning follows a structured and logical process when addressing complex problems. Typically, individuals begin by identifying the scope and nature of the problem, developing a comprehensive understanding of its context and underlying structure. Based on this understanding, they decompose the problem into smaller, manageable components and engage in step-by-step reasoning guided by prior knowledge and contextual cues. Eventually, they arrive at a well-informed judgment or conclusion through this analytical process.

Inspired by this human-like reasoning paradigm, we designed a standardized annotation format for harmful memes. This format aims to reflect the multi-step cognitive process humans use when identifying harmful content, thereby enhancing the model's interpretability and reasoning capability in a more human-aligned manner. To operationalize this idea, we break down each annotation into four clearly defined sections: **QUESTION**, **CAPTION**, **REASONING**, and **JUDGEMENT**. Each component mirrors a specific stage in the human decision-making process, from problem formulation to final conclusion. Below, we explain the purpose and structure of each section in detail:

QUESTION: To judge whether this meme is harmful, we need to analyze it based on the following five aspects: 1. Discrimination: Content that discriminates or provokes opposition based on race, gender, religion, nationality, physique, physical disabilities, LGBTQ+community, etc. 2. Offensive: Content that insults, belittles or maliciously attacks individuals or groups. In addition: Any content referencing sensitive events (e.g., disasters, pandemics, historical tragedies, racial issues, or violent incidents) in any form—humorous, sarcastic, dismissive, factual—is offensive. 3. Violence: Includes content that directly or indirectly encourages, depicts, or glorifies physical violence (e.g., fighting, weapons, gore), as well as verbal violence, such as the use of profanity, slurs, insults, or curses to attack or demean individuals or groups. 4. Vulgar: Content with sexual implications, obscenity, nudity, or crude language, including indirect expressions like puns, homophones, or symbolic use of objects and gestures to imply sexual acts or organs. Even without explicit language, content that clearly evokes sexual associations is considered vulgar. 5. Dissatisfaction: Content expressing strong discontent, resentment, or negativity toward individuals, groups, or society via aggressive satire, malicious sarcasm, or harmful exaggeration. Includes spreading pessimistic or nihilistic views without constructive intent. For each category, determine if it applies to the content. If a category does not apply, explicitly state 'Not applicable.'

CAPTION: Extract the meaning behind the text description, including background information, character information, homophonic information, and any other relevant context.

REASONING: Provide a step-by-step reasoning process based on the information extracted in the <CAPTION> section and your prior knowledge. For each of the five aspects listed in the <SUMMARY>,analyze whether the content fits into that category. Use logical explanations and examples to support your reasoning.

JUDGEMENT: Based on the information above to give your own judgment. The judgment must be strictly in the format: 'The image's label is harmful.' OR 'The image's label is harmless.'

B Data Annotation Examples

To further illustrate the annotation process and labeling criteria, this section provides several annotated examples of memes. As shown in Figure 5, the examples include both harmless and harmful cases. These instances clearly demonstrate how our annotation workflow interprets the meaning of different meme contents and conducts careful reasoning and analysis of their harmfulness based on five classification criteria.

In our annotation pipeline, GPT-4o acts as the main annotator. It is first given clear definitions of five harmful content categories: Discrimination, Offensiveness, Violence, Vulgar, and Dissatisfaction. Based on these definitions, the model starts by interpreting the meme in the **CAPTION** section—explaining its text, visuals, background, and any cultural or linguistic cues.

Next, in the **REASONING** section, the model analyzes whether the meme fits any of the harmful categories, using a step-by-step logic based on its understanding.

Finally, in the **JUDGEMENT** section, the model gives a clear decision: if the meme meets any harmful criteria, it is labeled as harmful; otherwise, it is harmless. This structured process reflects human-like reasoning and enhances the interpretability of the annotation results.

C Visual Demonstration of Results

To provide a clearer understanding of our model's performance and decision-making process, we present several visualized examples of annotated memes in Figure 6. These cases illustrate the model's binary classification of memes as either harmful or harmless, and demonstrate its Chain-of-Thought analysis through the stage of caption, reasoning and judgement.

First, in the caption stage, we analyze both the textual and visual components of the meme to extract relevant information, including sociocultural context, situational context, wordplay (e.g., homophones), and relationships between elements. Subsequently, during the reasoning phase, we conduct inference across five categories (Discrimination, Offensiveness, Violence, Vulgarity and Dissatisfaction) based on our predefined harmfulness detection criteria, leveraging the previously extracted meme information.

References

[1] Ali Anaissi, Junaid Akram, Kunal Chaturvedi, and Ali Braytee. 2025. Detecting and Understanding Hateful Contents in Memes Through Captioning and Visual Question-Answering. *arXiv preprint arXiv:2504.16723* (2025).

[2] Giovanni Burbi, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Mapping memes to words for multimodal hateful meme classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2832–2836.

[3] 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 ACM International Conference on Multimedia*. 5244–5252.

[4] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023. Prompting for multimodal hateful meme classification. *arXiv preprint arXiv:2302.04156* (2023).

REASONING:
1. Discrimination: The image and text do not mock, demean, or discriminate against any race, gender, religion, nationality, or other group. Thus, this aspect is Not applicable.
2. Offensive: The content does not insult or belittle anyone. The humor is simple and based on personal experiences. Therefore, this aspect is Not applicable.
3. Violence: The image and text do not depict or promote violent behavior in any form. Hence, this aspect is Not applicable.
4. Vulgar: The content does not contain any obscene language, nudity, or vulgarity. It is entirely appropriate. Thus, this aspect is Not applicable.
5. Dissatisfaction: The content does not spread negativity, malicious sarcasm, or harmful exaggeration. Instead, it offers light humor. Therefore, this aspect is Not applicable.

Meme text:
当我关灯后我的视力。过了一小会儿后我的视力。

JUDGEMENT:
The image's label is harmless.

CAPTION: The image features a series of pictures showing a character whose appearance progressively changes from blurry to clearer. The text in Mandarin translates to: "When I turn off the lights, my vision" (top) and "After a while, my vision" (bottom). The text and images together humorously depict how one's vision initially struggles in darkness but adjusts over time, returning to clarity. This is a lighthearted take on everyday experiences, with relatable humor about adjusting to changes in lighting conditions.

(a) Harmless

REASONING:
1. Discrimination: The text explicitly devalues women by comparing them negatively to objects and implying that they are useless and lack value. This is a clear case of gender-based discrimination.
2. Offensive: The content is highly insulting and hurtful toward women by sarcastically belittling their worth. This is offensive.
3. Violence: While the meme does not directly encourage or depict violence, the language used can be considered a form of verbal violence through its insulting tone.
4. Vulgar: Not applicable. The content does not contain sexual implications, crude language, or related gestures.
5. Dissatisfaction: The meme conveys a highly negative and harmful perspective toward women without any constructive intent, reflecting malicious sarcasm and harmful exaggeration.

Meme text:
DON'T TREAT WOMEN LIKE OBJECTS OBJECTS ARE USEFUL AND HAVE VALUE,UNLIKE WOMEN

JUDGEMENT:
The image's label is harmful.

CAPTION: The meme features a man in a retro art style, with text that reads, "DON'T TREAT WOMEN LIKE OBJECTS. OBJECTS ARE USEFUL AND HAVE VALUE, UNLIKE WOMEN." The text includes sarcastic, misogynistic content that devalues women while attempting to present itself in a humorous way. The underlying message seems to perpetuate gender discrimination.

(b) Harmful: Discrimination

REASONING:
1. Discrimination: Yes, the content associates an individual with a symbol of the KKK, which is historically tied to racial discrimination and violence. This imagery conveys discriminatory messaging.
2. Offensive: Yes, the content uses offensive historical symbolism tied to a hate group, specifically employing a KKK hood, which is inherently inflammatory and demeaning.
3. Violence: Indirectly applicable because the reference invokes the KKK, a group historically associated with violent behavior. While the image itself does not incite physical violence, it does allude to violent and harmful ideologies.
4. Vulgar: Not applicable. The content does not contain sexual innuendos, obscenities, or vulgar language.
5. Dissatisfaction: Yes, the content employs aggressive satire and malicious exaggeration critically targeting an individual by drawing comparisons to one of the most reviled hate groups in history, furthering negativity without constructive societal critique.

Meme text:
THANK YOU,MR.PRESIDENT,FOR WEARING YOUR MASK

JUDGEMENT:
The image's label is harmful.

CAPTION: The image depicts a cartoon illustrating a political debate. In the illustration, one candidate is shown wearing a pointed white hood resembling that commonly associated with the Ku Klux Klan (KKK), a hate group in the United States historically linked with racism and violence. The text reads, "Thank you, Mr. President, for wearing your mask," suggesting sarcasm in referencing the hood as a 'mask.' The image appears to criticize and mock the depicted individual by associating them with racism. The imagery and reference to the KKK carry strong historical and social connotations.

(c) Harmful: Dissatisfaction

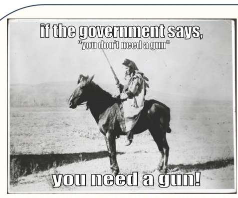
REASONING:
1. Discrimination: Not applicable. The content does not discriminate or target any specific group based on race, gender, religion, etc.
2. Offensive: Not applicable. There's no indication of malicious attacks, insults, or reference to sensitive events.
3. Violence: Not applicable. No direct or indirect depiction or promotion of violence is present in the image.
4. Vulgar: Applicable. The content heavily implies sexual innuendos through the actions and attire of the characters, as well as the deliberate inclusion of the numeric code, likely referencing adult material. This constitutes vulgarity due to the sexual associations evoked.
5. Dissatisfaction: Not applicable. The image does not convey aggressive societal negativity, discontent, or harmful sarcasm.

Meme text:
現在要幫您檢查喔

JUDGEMENT:
The image's label is harmful.

CAPTION: The text in the image translates to "Now we are going to examine you" with two women dressed as nurses interacting closely with a man in what appears to be a suggestive context. Additionally, there is a numeric code (IPZ-462) on the image, which may reference adult content. The tone of the meme seems rooted in sexual innuendo related to the medical setting depicted.

(d) Harmful: Vulgar



REASONING:
1. Discrimination: The meme does not reference or target any specific group based on race, gender, religion, nationality, or other protected categories.
2. Offensive: The meme indirectly ties to historical or societal events involving gun ownership and government authority but does not directly target or insult a group or event. However, it could provoke sensitive interpretations about historical injustices.
3. Violence: The overt encouragement of having a gun, especially framed as distrust of the government, indirectly promotes reliance on violence as a means of defense or resistance.
4. Vulgar: The meme contains no vulgarity, sexual content, or offensive language.
5. Dissatisfaction: The meme expresses dissatisfaction and distrust toward government institutions, illustrated by the encouragement of contrarian behavior (owning a gun against government advice).

Meme text:
if the government says, "you don't need a gun" you need a gun!

JUDGEMENT:
The image's label is harmful.

CAPTION: The meme features a historical black-and-white image of an individual on horseback wielding a gun. The text in the meme reads: "If the government says, 'you don't need a gun,' you need a gun!" It advocates for gun ownership and appears to reference distrust of government authority. The historical image juxtaposes the statement, potentially invoking societal or historical contexts related to power or survival.

(e) Harmful: Violence

REASONING:
1. Discrimination: Not applicable. The image does not target any specific group or express discriminatory ideas based on race, gender, religion, etc.
2. Offensive: Applicable. The use of profanity ("CNM") serves as a direct insult ("f*** your mother"), which could be perceived as offensive and degrading to some viewers. The humor derived from profanity can be taken as disrespectful.
3. Violence: Not applicable. The meme does not depict or promote physical or verbal violence beyond the use of profane language.
4. Vulgar: Applicable. The use of "CNM" in the image directly incorporates crude language, which is inherently vulgar.
5. Dissatisfaction: Not applicable. The image does not express dissatisfaction, resentment, or negativity toward a specific person, group, or society. Its intent appears to be provocative humor rather than an expression of discontent.

Meme text:
CNM 比个心情

JUDGEMENT:
The image's label is harmful.

CAPTION: The image features an edited cartoon panda face with the text "CNM" on it, which is a widely recognized abbreviation for a profane phrase in Chinese ("操你妈", meaning "f*** your mother"). Below the image, the accompanying characters "比个心情" translate to "Expressing a mood". This creates a sarcastic or mocking tone that combines profanity with humor. The image is intended to amuse by using crude language.

(f) Harmful: Offensive

Figure 5: Illustrates several annotated examples from our MemeMind dataset, covering both harmless and harmful cases, as well as different types of harmful content. These include: (b) Discrimination, (c) Dissatisfaction, (d) Vulgar, (e) Violence, (f) Offensive.

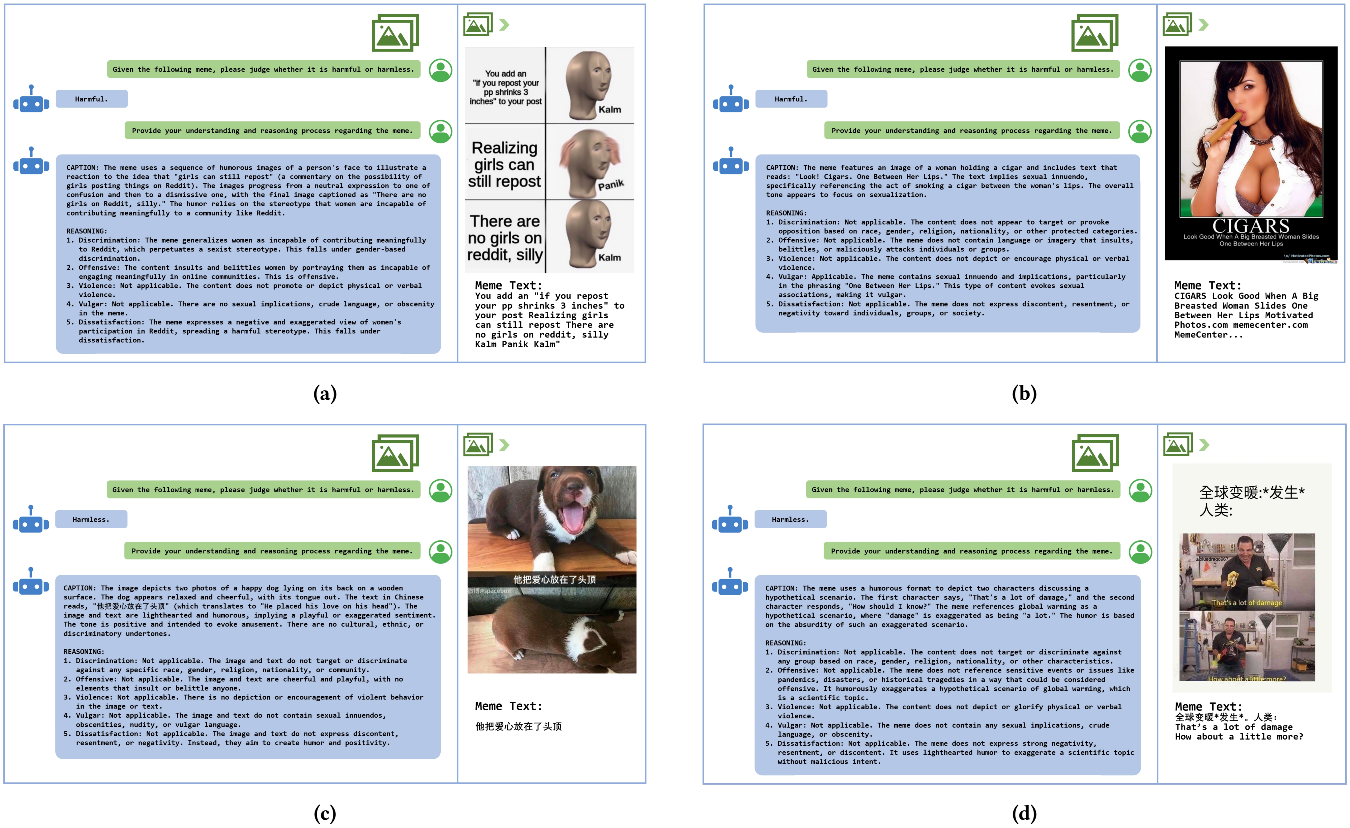


Figure 6: Visualization examples of the model's output on our benchmark dataset. The selected cases include both harmful (a, b) and harmless (c, d) memes, encompassing content in both Chinese and English. These examples demonstrate the model's ability to detect the harmfulness of memes in multilingual and multicultural contexts, as well as its capacity to understand and reason about their content.

[5] Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018. Cross-dataset adaptation for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5716–5725.

[6] Julia R DeCook. 2018. Memes and symbolic violence:# proudboys and the use of memes for propaganda and the construction of collective identity. *Learning, Media and Technology* 43, 4 (2018), 485–504.

[7] Katie M Duchscherer and John F Dovidio. 2016. When memes are mean: Appraisals of and objections to stereotypic memes. *Translational Issues in Psychological Science* 2, 3 (2016), 335.

[8] 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 International Workshop on Semantic Evaluation (SemEval-2022)*. 533–549.

[9] Shlok Gilda, Luiz Giovanini, Mirela Silva, and Daniela Oliveira. 2022. Predicting different types of subtle toxicity in unhealthy online conversations. *arXiv preprint arXiv:2106.03952* (2022).

[10] Biagio Grasso, Valerio La Gatta, Vincenzo Moscato, and Giancarlo Sperlì. 2024. KERMIT: knowledge-empowered model in harmful meme detection. *Information Fusion* 106 (2024), 102269.

[11] Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM web conference 2022*. 3651–3655.

[12] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409* (2017).

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2022).

[14] Jianzhao Huang, Hongzhan Lin, Ziyan Liu, Ziyang Luo, Guang Chen, and Jing Ma. 2024. Towards Low-Resource Harmful Meme Detection with LMM Agents. *arXiv preprint arXiv:2411.05383* (2024).

[15] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[16] 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. *arXiv preprint arXiv:2005.04790* (2020).

[17] Girish A Koushik, Diptesh Kanojia, Helen Treharne, and Aditya Joshi. 2025. CAMU: Context Augmentation for Meme Understanding. *arXiv preprint arXiv:2504.17902* (2025).

[18] Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. *arXiv preprint arXiv:2210.05916* (2022).

[19] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

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

[21] Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of chinese harmful memes. *arXiv preprint arXiv:2410.02378* (2024).

[22] Susanna Paasonen. 2019. Online pornography. *The SAGE handbook of web history* (2019), 551–563.

[23] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework

for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184* (2021).

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

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

[26] Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. Memeclip: Leveraging clip representations for multimodal meme classification. *arXiv preprint arXiv:2409.14703* (2024).

[27] Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022. DISARM: Detecting the victims targeted by harmful memes. *arXiv preprint arXiv:2205.05738* (2022).

[28] Shivam Sharma, Firoj Alam, Md Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274* (2022).

[29] Limor Shifman. 2013. *Memes in digital culture*. MIT press.

[30] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[31] Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the international ACM SIGIR conference on research and development in information retrieval*. 2887–2899.

[32] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440* (2024).

[33] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12104–12113.

[34] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530* (2016).