Towards Comprehensive Detection of Chinese Harmful Memes

Junyu Lu[1], Bo Xu[1], Xiaokun Zhang[1], Hongbo Wang[1],
Haohao Zhu[1], Dongyu Zhang[2], Liang Yang[1,3], Hongfei Lin[1]*
[1] School of Computer Science and Technology, [2] School of Foreign Languages,
[3] Key Laboratory of Social Computing and Cognitive Intelligence,
Dalian University of Technology, China
dutljy,zhuhh@mail.dlut.edu.cn, dawnkun1993,1846742523a@gmail.com
xubo,zhangdongyu,liang,hflin@dlut.edu.cn

Abstract

Harmful memes have proliferated on the Chinese Internet, while research on detecting Chinese harmful memes significantly lags behind due to the absence of reliable datasets and effective detectors. To this end, we focus on the comprehensive detection of Chinese harmful memes. We construct TOXICN MM, the first Chinese harmful meme dataset, which consists of 12,000 samples with fine-grained annotations for various meme types. Additionally, we propose a baseline detector, Multimodal Knowledge Enhancement (MKE), incorporating contextual information of meme content generated by the LLM to enhance the understanding of Chinese memes. During the evaluation phase, we conduct extensive quantitative experiments and qualitative analyses on multiple baselines, including LLMs and our MKE. The experimental results indicate that detecting Chinese harmful memes is challenging for existing models while demonstrating the effectiveness of MKE.[1]

***Disclaimer**: The samples presented by this paper may be considered offensive.*

1 Introduction

With the development of the Internet, harmful memes on the web have become increasingly rampant. Harmful memes are typically defined as multimodal units consisting of an image and embedded text that cause harm to an individual, an organization, a community, or a social group by specifically targeting social entities (Pramanick et al., 2021a; Sharma et al., 2022b). They may exacerbate social divisions, trigger discriminatory behaviors, and harm social harmony and unity (Kiela et al., 2020). Due to their negative impact on society, the widespread dissemination of harmful memes has been widely recognized as a growing concern.

In recent years, researchers have made substantial progress in detecting harmful memes. Several datasets, including HMC (Kiela et al., 2020), MMHS (Gomez et al., 2020), Harm-C, and Harm-P (Pramanick et al., 2021a), have been established, and various detectors have been proposed Hee et al. (2022); Aggarwal et al. (2023); Pramanick et al. (2021b); Sharma et al. (2022a); Cao et al. (2022). However, most existing studies only focus on English memes. In contrast, Chinese harmful meme detection remains largely unexplored, presenting challenges in building reliable datasets and developing effective detectors.

On one hand, the types of Chinese harmful memes are diverse. In addition to those targeting specific social entities, many memes on Chinese platforms contain general offense, sexual innuendo, or

*Corresponding author

[1]Resources of this paper are available at https://github.com/DUT-lujunyu/ToxiCN_MM.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

Figure 1: Illustration of memes in Chinese. (a) is a harmless meme that humorously expresses concern about math homework. (b) is a targeted harmful meme conveying gender bias through the height differences between the man and woman. (c) contains general offense without specific targets, where "*vegetable dog*" is a Chinese insult, implying incompetent people. (d) subtly conveys sexual innuendo with an idiom "*shy things*" (alluding to "*sexual intercourse*"). (e) spreads dispirited culture by comparing oneself to "*garbage*".

dispirited culture (Liu and Xu, 2016), as shown in Figure 1. Despite lacking specific targets, they still exhibit potential toxicity, subtly propagating negative values that could lead to serious consequences like violent acts and sexual harassment (Lin and Zhang, 2019). To adapt to the online environment, it is crucial to consider this diversity when constructing the dataset.

On the other hand, understanding the semantics of Chinese harmful memes presents a significant challenge for detectors, necessitating contextual information from both textual and visual elements. For example, Exp. (a) of Figure 1 illustrates gender bias through the height difference between the man and the woman in the image. In contrast, the inline text of Exp. (b) introduces a Chinese insult, *"vegetable dog"*, to tease others, which implies incompetent people. Therefore, incorporating such information is necessary for the effective detection of Chinese harmful memes.

In this paper, we facilitate the detection of Chinese harmful memes, primarily focusing on two aspects: **dataset construction** and **detector development**. For the dataset construction, we first propose the definition of "*Chinese harmful memes*" as guidance, accurately adapting to the Chinese online environment. Based on the definition, we focus on both targeted harmful memes and those exhibiting potential toxicity without specific targets. We conduct fine-grained annotation for harmful memes collected from Chinese online platforms, analyzing their harmful types and combination features of textual and visual information. TOXICN MM dataset is then constructed, encompassing 12,000 samples containing different harmful types. Two progressive tasks are established using the dataset: (I) Detect if a meme is harmful, and (II) If harmful, further identify its harmful type.

For the detector development, we present a Multimodal Knowledge Enhancement (MKE) detector, integrating the contextual information of meme content for effective detection. We first utilize the large language model (LLM) to capture the context of both the text and image of the meme, leveraging its extensive knowledge acquired through pre-training Zhao et al. (2023). This information is then integrated into a trainable detector as enhanced captions to improve the understanding of memes. In the experimental phase, we evaluate the detection performance of various baselines, including both traditional pre-trained language models (PLMs) and large language models (LLMs), providing a benchmark for evaluation. Experimental results show the effectiveness of our MKE. The main contributions of this paper are summarized as follows:

- We construct, to the best of our knowledge, the first Chinese harmful meme dataset TOXICN MM, which comprises 12,000 diverse samples. We conduct a fine-grained annotation to analyze their harmful types and modality combination features.
- We present Multimodal Knowledge Enhancement (MKE) as a baseline detector, integrating contextual information of meme content generated by the LLM to improve the detector's understanding of Chinese harmful memes.
- We utilize TOXICN MM as a benchmark to evaluate the detection performance of various baseline models. Extensive quantitative experiments and qualitative analysis illustrate the effect of MKE. We summarize the challenges of Chinese harmful memes detection.

Work	Language	Size	Ratio	Pot. Tox.
HarMeme (Pramanick et al., 2021a)	English	3,544	34.96%	✗
Harm-C (Pramanick et al., 2021b)	English	3,013	35.31%	✗
Harm-P (Pramanick et al., 2021b)	English	3,020	49.21%	✗
HMC (Kiela et al., 2020)	English	10,000	38.00%	✗
MMHS150K (Gomez et al., 2020)	English	150,000	24.68%	✗
TamilMemes (Suryawanshi et al., 2020)	Indian	2,969	65.71%	✗
MUTE (Hossain et al., 2022)	Bengali	4,158	37.87%	✗
MAMI (Fersini et al., 2022)	English	11,000	50.00%	✗
TOXICN MM (ours)	Chinese	12,000	31.89%	✓

Table 1: Information of harmful meme datasets, in terms of *Language*, *Size*, *Ratio* of harmful samples, and whether containing memes exhibiting potential toxicity (*Pot. Tox.*) without specific targets.

2 Related Work

Harmful Meme. In recent years, researchers have noticed the importance of detecting harmful memes. Several datasets have been established Kiela et al. (2020); Gomez et al. (2020); Pramanick et al. (2021a). Nevertheless, most current studies only focus on English, while research on detecting Chinese harmful memes remains unexplored. To this end, we present the first Chinese harmful meme dataset TOXICN MM. Here we list Table 1 to compare existing datasets with TOXICN MM.

Existing studies define "*harmful memes*" as memes that cause harm to specific social entities, based on their social attributes such as religion, race, and gender (Pramanick et al., 2021a). However, this definition does not fully apply to the Chinese Internet, where harmful memes often exhibit potential toxicity without specific targets but still perpetuate negative cultural values (Liu and Xu, 2016; Zhang and Zhao, 2021). In this paper, we consider different harmful types for comprehensive detection.

Several methods have been proposed for harmful meme detection, primarily focusing on modeling based on targets of memes Pramanick et al. (2021b); Sharma et al. (2022a); Koutlis et al. (2023); Ji et al. (2023). However, they do not apply to Chinese harmful memes, which often lack specific targets. Despite this limitation, some methods enhance the detector's understanding of memes by integrating image descriptions to make decisions Wang et al. (2024b); Cao et al. (2022). These studies still inspire us to integrate more comprehensive contextual information of meme content into detectors.

Toxic Language. Harmful memes are closely linked to toxic language (Kiela et al., 2020), which are rude, disrespectful, or unreasonable, and can drive people away from conversations (Dixon et al., 2018). Chinese harmful memes frequently feature toxic language in their inline text, utilizing slang and linguistic phenomena like homophony. Therefore, understanding Chinese memes necessitates the incorporation of linguistic knowledge.

Additionally, labeling both harmful memes and toxic language is often subjective. Several studies have addressed this issue by focusing on mitigating the subjective bias of annotators during the construction of toxic language datasets (Waseem and Hovy, 2016; Ross et al., 2016; Zeinert et al., 2021; Lu et al., 2023; Wang et al., 2023, 2024a), enhancing the reliability of datasets. In this paper, we adopt these measures as a reference in the annotation process of our TOXICN MM.

3 Dataset Construction

3.1 Overview

In this section, we detail the construction of our TOXICN MM dataset. We first define "Chinese harmful memes" to guide the dataset annotation. We then conduct fine-grained annotation for memes collected from Chinese platforms. In addition to the basic binary labels, we analyze harmful memes from both harmful types and combinations of inline text and image information. After the annotation, statistics of TOXICN MM are presented. The diagram of dataset construction is shown in Figure 2.

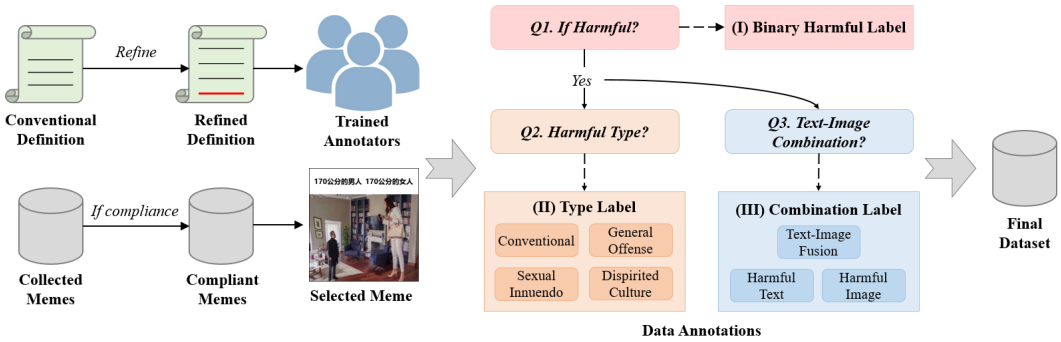

Figure 2: Illustration of TOXICN MM construction procedure. According to the definition (Section 3.2), data collection and filtering (Section 3.3) and fine-grained annotations (Section 3.4) are conducted sequentially. The final statistics of TOXICN MM are presented in Section 3.5.

3.2 Definition Development

The recognized definition of "*harmful meme*" typically pertains to memes that target specific social entities. However, numerous memes on the Chinese Internet diverge from this definition by only propagating negative values without specific targets, which can be equally detrimental to society. To adapt to the Chinese online environment, a refined definition is necessary. Here we introduce the definition of Chinese harmful memes:

Chinese harmful memes are multimodal units consisting of an image and Chinese inline text that have the potential to cause harm to an individual, an organization, a community, a social group, or society as a whole. These memes can range from offense or joking that perpetuate harmful stereotypes towards specific social entities, to memes that are more subtle and general but still have the potential to cause harm. It is important to note that Chinese harmful memes can be created and spread intentionally or unintentionally. They often reflect and reinforce underlying negative values and cultural attitudes on the Chinese Internet, which are detrimental from legal or moral perspectives.

According to the definition, we further identified the most common harmful types of memes on Chinese platforms based on the consensus of social psychology (Liu and Xu, 2016; Lin and Zhang, 2019) and communication (Peng, 2019; Zheng, 2016) studies. Specifically, it mainly includes *targeted harmful*, *general offense*, *sexual innuendo*, and *dispirited culture*. The harm of these memes to individuals and society has been widely discussed. In this study, we focus on these harmful types when constructing the dataset.

3.3 Data Collection and Filtering

Data collection is the basic work of constructing datasets, and its breadth and quality greatly affect the subsequent research. To ensure a comprehensive dataset, we collect Chinese memes from two well-known public online platforms, *Weibo* and *Baidu Tieba*, both widely representative of local users in China with active meme communities. We first conduct a random crawl to obtain a diverse set of memes. To maximize the inclusion of harmful memes, we further focus our data crawl on sensitive topics commonly debated online (*e.g.*, "*gender*" and "*region*"). Additionally, we also target memes expressing negative emotions and attitudes (*e.g.*, "*crazy*" and "*Dispirited Culture*") to enrich the dataset with samples potentially exhibiting toxicity. A total of about 14k memes are collected. We then de-duplicate the data and filter out dirty samples including unreadable memes. The final dataset contains 12k refined memes.

Subsequently, we utilize `Baidu-OCR` to extract inline text from memes, which offers a high-precision service for Chinese text recognition. To further enhance the sample quality, we also introduce a manual review process to examine the accuracy of the extracted text. Specifically, we normalize the text by adding appropriate separators and removing additional line breaks and spaces.

3.4 Data Annotation

3.4.1 Annotator Selection and Training

Before the formal annotation process, it is crucial to select annotators carefully and mitigate their subjective bias, as this can significantly impact the quality of the dataset (Waseem and Hovy, 2016). To this end, we adopt the following measures: The majority of active users on Chinese platforms are between 12 and 35 years old. Considering Chinese laws that restrict individuals under 18 from engaging in activities that could harm their physical or mental health, we selected annotators aged 18 to 35. We assessed the annotators' proficiency in Chinese meme culture through questionnaires and ensured diversity in terms of gender, region, and education level to enhance reproducibility. The demographics of annotators are shown in Table 2.

Characteristic	Demographics
Gender	Male: 5, Female: 5
Age	<20: 3, 20~30: 5, >30: 2
Race & Region	Asians: 7, Others: 3
Education	UG: 3, PG: 4, PhD: 3

Table 2: Annotators demographics.

During the training of annotators, we provided definitions of Chinese harmful memes, their various types, and conducted case analyses with diverse examples. To evaluate the annotators' abilities, we introduced three test groups of 100 memes each. Annotators labeled the memes independently, with researchers finalizing the labels. Post-round discussions were held to reduce errors, and detailed criteria, including edge cases, were established. Annotators improved from 63% accuracy in the first round to 78% in the final round, demonstrating the effectiveness of the training.

3.4.2 Label Annotation

To guarantee the consistency of label annotation, we establish a comprehensive annotation framework as a guideline. The specific process includes the following three stages.

Whether Harmful. The foundation of the labeling is to determine whether a meme is harmful or benign, which is a binary annotation. We strictly follow the definition of "*Chinese harmful memes*", focusing not only on targeted harmful memes but also on samples exhibiting potential toxicity without specific targets.

Harmful Type. In the second stage, we further refine the categorization of harmful memes, including targeted harmful, general offensive, sexual innuendo, and dispirited culture. The annotation criteria for each harmful type are provided below. **Targeted Harmful** memes express disgust, prejudice, or stereotypes towards specific individuals or social groups. In contrast, **General Offensive** memes encompass sarcastic or rude content but lack specific targets. We also adhere to psychological and sociological definitions to classify the other two types: **Sexual Innuendo** refers to memes that imply sexual intent to provoke sexual arousal (Bell, 1997). Here we label memes that contain suggestive elements but not sexism or sexual assault as such samples to distinguish them from targeted harmful memes. And **Dispirited Culture** is characterized by the integration of decadent and desperate emotions, conveying a self-negative attitude (Dong et al., 2017).

Modality Combination. As multimodal units, harmful memes consist of both textual and visual modality, expressing toxicity through fused or independent features (Kiela et al., 2020). To gain a more comprehensive understanding of how harmful content is propagated via memes, we classify them based on the toxic manifestation of these two modalities, exploring their individual and combined effects. Among them, **Text-Image Fusion** memes exhibit toxicity only through the combined effect of both modalities, while the text and image separately remain benign. In contrast, **Harmful Text** and **Harmful Image** categories refer to one modality (either the text or the image) that independently exhibits toxicity.

During the label annotation, each meme is labeled by at least three annotators. And we use a majority vote to assign the final label. In addition, specific targets of targeted harmful memes are provided. We then discuss the Inter-Annotator Agreement (IAA) of each granularity, as shown in Appendix B.3.

3.5 Statistics Description

For subsequent model training and evaluation, all samples in TOXICN MM are divided into a training set and a test set at a ratio of 8:2, as detailed in Table 3. We note that there exists a sample imbalance

Split	N-Harm.	Harm.	Harmful Type Category				Combination Category			Total
			Tg.	Off.	Sex.	Disp.	T-I	Harm.T	Harm.I	
Train	6,538	3,062	813	1,198	731	320	1,082	1,754	276	9,600
Test	1,635	765	203	300	183	79	271	438	69	2,400
Total	8,173	3,827	1,016	1,498	914	399	1,353	2,192	345	12,000

Table 3: Basic statistics of TOXICN MM, listing the number of non-harmful (*N-Harm.*) and harmful (*Harm.*) samples, containing targeted harmful memes (*Tg.*), general offense (*Off.*), sexual innuendo (*Sex.*), and dispirited culture (*Disp.*), as well as each modality combination category (including text-image fusion (*T-I*), harmful text (*Harm.T*) and harmful image (*Harm.I*)).

among different categories of harmful samples. Memes containing general offensive content (*Off.*) constitute nearly 40% of harmful memes. Regarding modality combinations, over 50% of the inline text of harmful memes is harmful. Given that the data distribution accurately reflects the real state of platforms, we do not introduce supplementary sampling to address existing imbalances.

We then analyze the modality combinations across different harmful types, as shown in Table 4. Each type displays distinct patterns in its modality combinations. For example, memes containing general offense or dispirited culture (*Disp.*) mainly feature inherently harmful inline text. In contrast, over 50% of targeted harmful memes (*Conv.*) integrate multimodal features to express toxicity, where both text and image are individually benign. Moreover, there are 63 memes in which both text and image exhibit toxicity.

	T-I	Harm.T	Harm.I	Total
Tg.	575	404	47	1,016
Off.	198	1,247	93	1,498
Sex.	431	307	186	914
Disp.	149	234	19	399
Total	1,353	2,192	345	12,000

Table 4: Modality combination distribution of different harmful types in TOXICN MM.

4 Detector Development

4.1 Overview

To improve the detector's understanding of memes, we present a baseline detector, Multimodal Knowledge Enhancement (MKE), which integrates contextual information of meme content for more accurate predictions. We first leverage the LLM to capture the contextual context of memes and generate enhanced captions. Then, we fine-tune the detector by integrating the original inputs (i.e., text-image pairs) with the generated captions. The illustration of MKE is shown in Figure 3.

For a given meme, its inline text and image are encoded by modality-specific encoders, represented as $S \in \mathbb{R}^{d_s}$ for text and $V \in \mathbb{R}^{d_v}$ for the image, where d_s and d_v denote the dimensions of the textual and visual vector spaces, respectively.

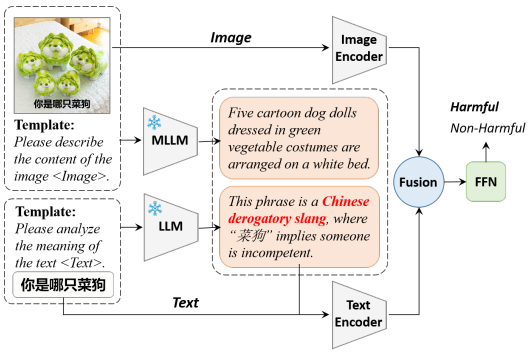

Figure 3: Overview of MKE. The translation of the inline text is "what vegetable dog are you".

4.2 Knowledge Mining

We instruct the LLM to generate enhanced captions for the meme by designing the instruction template, which respectively captures the contextual information from both the inline text and image. To improve the understanding of the inline text that may contain slang, we enable LLM to incorporate language features unique to Chinese for semantic analysis. The template is as follows: "*Considering Chinese linguistic characteristics, please analyze the meaning of the text <Text>.*" We further convert the image into textual descriptions with the multimodal large language model (MLLM), capturing

harmful elements in the context of Chinese culture. The template is designed as "*Considering Chinese cultural background, please describe the content of the image <Image>.*"

In the process of knowledge mining, all parameters of LLMs are frozen. To facilitate knowledge integration for subsequent detectors, the captions of inline text and images are represented by the text encoder, denoted as K_s and $K_v \in \mathbb{R}^{d_s}$.

4.3 Knowledge Integration

To leverage contextual information, we employ a cross-attention mechanism to integrate the inline text with two types of caption information, due to the consistency of the textual vector space. The feature introducing the textual caption K_s is defined as $S_{K_s} = \mathrm{Softmax}\left(S{K_s}^T/\sqrt{d_s}\right)S$. Similarly, the feature introducing the visual caption K_v is obtained and denoted as S_{K_v}. We then incorporate these features into a knowledge-enhanced representation $S_K = \mathrm{Mean}(S, S_{K_s}, S_{K_v})$, where $S_K \in \mathbb{R}^{d_s}$. Next, we concatenate S_K with the original image feature V to obtain the final representation of a meme, denoted as $C \in \mathbb{R}^{d_c}$, where $d_c = d_s + d_v$. C is then processed by a trainable classifier, which applies a linear transformation followed by a softmax function to produce detection probabilities.

5 Experiments

5.1 Tasks and Baselines

We utilize TOXICN MM as the benchmark for Chinese harmful meme detection. Specifically, we establish two progressive tasks. (I) **Harmful Meme Detection**, a binary classification task, detects if a meme is harmful; (II) **Harmful Type Identification**, a multi-classification, further identifies its harmful type, including targeted harmful memes, general offense, sexual innuendo, or dispirited culture. In addition to MKE, we evaluate the performance of various baselines, including both unimodal and multimodal models. For unimodal models, we utilize RoBERTa (Liu et al., 2019), GPT-3.5, and GPT-4 (text input) as text-only models, while ResNet (He et al., 2016) and ViT (Dosovitskiy et al., 2021) serve as image-only models. For multimodal models, we employ CLIP (Radford et al., 2021), the fusion of RoBERTa and ViT, which concatenates representations of the text and image for classification, and GPT-4 (text and image input).

5.2 Implementation

We adopt precision (*P*), recall (*R*), and macro F_1-score (F_1) as metrics. We also report the F_1 of harmful memes and each harmful type. We respectively utilize CLIP and the fusion of RoBERTa and ViT as the backbones of MKE, and we use GPT-4 to generate enhanced captions. For conventional PLMs, we fine-tune their parameters and select the best-performing model based on test set outcomes. For LLMs, we evaluate their performance in a zero-shot scenario, using instruction templates in Chinese. More details are provided in Appendix B.5.

5.3 Results and Discussions

In this section, we present our experimental results and conduct a detailed analysis. The performance of baselines is evaluated across two tasks, as shown in Table 5. From the results, we can observe that:

(1) In contrast to LLMs, conventional fine-tuned pre-trained baselines (i.e., CLIP and the combination of RoBERTa and ViT) achieve better detection performance, indicating their effectiveness in specific tasks. When considering the modality of input information, RoBERTa, which solely utilizes the inline text of memes, achieves a significantly higher F_1 score (average increase of 8.4%) than vision-based methods such as ResNet and ViT, which solely utilize images. This result supports the conclusion drawn in (Hee et al., 2022), namely that text comprehension plays a more crucial role than image understanding in the detection of harmful memes.

(2) GPT-4 and GPT-3.5 show similar performance in binary *harmful meme detection* when only the inline text is provided, and there is a clear enhancement in the multiclass *harmful type identification* task. After incorporating the image input, GPT-4 shows the best detection performance for sexual innuendo (*Sex.*) memes, while its performance decreases for general offense (*Off.*) and dispirited culture (*Disp.*). Referring to Table 4, we observe that most samples of *Sex.* exhibit toxicity through

Modality	Model	Harmful Meme Detection				Harmful Type Identification						
		P	R	F1	$F1_{Harm.}$	P	R	F1	$F1_{Tg.}$	$F1_{Off.}$	$F1_{Sex.}$	$F1_{Disp.}$
Text	GPT3.5	69.46	66.59	67.46	53.25	36.93	32.60	32.45	27.42	29.01	10.52	14.63
	GPT4	74.52	65.59	68.01	51.78	58.29	41.81	44.86	11.43	62.22	32.26	34.15
	RoBERTa	75.52	77.54	76.36	66.48	53.24	60.06	55.85	48.79	71.81	42.70	29.23
Image	ResNet	66.61	66.92	66.76	53.76	35.66	36.23	36.46	16.67	51.28	20.93	12.28
	ViT	68.97	68.61	68.78	57.24	43.38	37.86	39.10	24.17	54.57	34.32	9.01
Multimodal	GPT4	74.67	68.64	70.11	55.77	58.87	41.77	43.89	23.53	32.56	55.74	23.53
	Fusion	77.77	79.18	78.39	69.61	58.93	60.58	59.35	50.24	73.35	47.85	38.71
	+ K_s	78.17	79.32	79.04	70.33	60.09	63.38	61.28	51.02	75.60	48.75	**43.06**
	+ K_v	77.93	79.32	78.55	69.85	59.16	60.67	59.61	51.41	75.00	48.68	36.23
	+ MKE	77.96	80.96	79.16	70.22	**62.41**	62.08	**62.17**	53.27	74.54	56.10	39.24
	CLIP	78.95	80.26	79.54	71.28	54.85	64.95	57.85	49.58	74.65	49.64	26.92
	+ K_s	79.23	80.71	79.89	71.72	56.24	**66.45**	59.23	47.80	77.17	54.72	27.78
	+ K_v	79.39	**80.93**	80.07	71.96	57.20	64.27	59.24	**53.41**	76.97	52.53	25.24
	+ MKE	**79.76**	80.79	**80.23**	**72.35**	60.38	63.52	61.47	51.52	**77.19**	**57.14**	33.33

Table 5: Detection performance of baselines. Results show the mean of *P*, *R*, and macro F_1, where the **bold** and underline scores respectively represent the optimal and suboptimal values. *Fusion* refers to the fusion of RoBERTa and ViT, and K_s and K_v respectively denote introducing enhanced captions of inline text and image. All results are statistically significant, as determined by a t-test ($p < 0.01$).

image-text fusion or harmful images, whereas images of *Off.* and *Disp.* are mostly benign. This suggests that the toxicity of visual information has a significant impact on the decisions of GPT-4. We will further explain this in the following case study.

(3) Our MKE demonstrates superior performance, with an average macro-F1 -score improvement of 0.73% and 3.22% over the backbone models for both tasks. This improvement illustrates the effectiveness of introducing contextual information of meme content for detecting Chinese harmful memes. Ablation studies show that both enhanced captions for inline text and images contribute to the detector's deeper understanding of memes, leading to more precise classifications. Additionally, the degree of performance enhancement varies depending on the type of harmful meme. For instance, for targeted harmful memes (*Tg.*), where toxicity often relies on the combination of image and text, image captions provide a greater boost (2.07%). In contrast, for memes expressing dispirited culture (*Disp.*), where the text is typically harmful, inline text captions lead to a larger improvement (2.58%).

5.4 Case Study

To further illustrate the rationales of MKE, we provide several case studies, as shown in Table 6. We list enhanced captions of harmful memes and the predictions of other models for reference. We also instruct GPT-4 to generate reasons for its detection decisions. We do not introduce additional templates to standardize its reasoning to reflect GPT-4's true understanding of memes more accurately.

Exp. (a) is a targeted harmful meme towards Asians. Through the caption, we observe that GPT-4 understands the meme's meaning solely through the inline text, recognizing the high standard of Asian parents on their children's academic performance. This highlights GPT-4's strong contextual understanding. After integrating this information, compared to the backbone (Fusion), our MKE model makes the correct decision, illustrating that incorporating contextual information of meme content enhances the model's understanding of memes.

For more insight into the challenges of detecting Chinese harmful memes, we manually inspected the samples misclassified by most baselines. Two main types of errors are summarized.

Type I error: Benign information contained in harmful memes can influence the judgment of models, resulting in incorrect detection. In Exp. (b), when presented solely with the inline text, GPT-4 accurately interprets the meme's meaning, comparing "*I*" with a "*little mouse*" to convey a dispirited culture. However, upon introducing the image, GPT-4 mistakenly interprets the mouse as being "*gently stroked*" and incorrectly categorizes the meme as harmless. This suggests that the model may overlook the potential toxicity of memes due to the seemingly benign nature of a certain modal. In contrast, MKE integrates original sample and caption information to make the correct judgment.

	(a)	(b)	(c)
Harmful Meme	*When Asian parents find out their children are learning C instead of A+*	*Life has finally come down hard on me, a little mouse.*	*What are you dog-bananaing at?*
Category	Targeted Harmful, Text-Image Fusion	Dispirited Culture, Harmful Text	General Offensive, Harmful Text
Text Caption	Asian parents demand high academic performance, expecting their children to get excellence (A+) and not average grades (C).	It means that life is stressful or difficult. It implies that one is as small or unimportant as a mouse, involving dispirited culture.	"狗蕉" may be a slang term, and the text meaning may be jokingly describing some combination of dog and banana.
Image Caption	Four grieving Asian adults were crying and holding each other up.	One man gently cradled a small hamster-like animal in his hand.	The image shows a yellow banana with one end replaced with a pattern of a dog's head.
Explanation	This meme can be considered a humorous reference to cultural stereotypes, exaggerating Asian parents' concern for their children's academic performance.	The meaning of this meme is to make a self-joke about the hardships of life. It is just a humorous way to express someone's helpless sense of life without toxicity.	This meme is just a kind of humor. By combining the dog and banana together with words, it produces a humorous effect that subverts expectations.
Prediction	GPT-4 (only text): ✓, GPT-4: ✓, CLIP: ✓, Fusion: ✗, MKE (Fusion): ✓.	GPT-4 (only text): ✓, GPT-4: ✗, CLIP: ✗, Fusion: ✗, MKE (Fusion): ✓.	GPT-4 (only text): ✗, GPT-4: ✗, CLIP: ✗, Fusion: ✗, MKE (Fusion): ✗.

Table 6: Illustration of case study. Highlighted in green and red within the decision reasons are the accurate implications of toxicity and misinformation. ✓ and ✗ represent the success and failure of model prediction. In the implementation, all descriptions are provided in Chinese.

Type II error: Harmful memes containing unique cultural backgrounds are easily missed by models. Among them, the most challenging samples are the memes whose inline text contains specific expressions in Chinese, such as homophones and metaphors. Take Exp. (c) for example, where the term "*dog-banana*" is a homonym of "*bark*" in Chinese. Therefore, this meme is essentially a harmful meme containing general offense, implicitly expressing dissatisfaction with another person. Due to the lack of related knowledge, current models fail to comprehend the underlying semantics of these memes, making them difficult to detect successfully.

These case studies further illustrate that Chinese harmful memes detection is a complex multimodal semantic understanding task, which is challenging for existing models. The error analysis shows that fully integrating image and text information and introducing more comprehensive knowledge about Chinese culture are both crucial for effectively detecting harmful memes.

6 Conclusions and Future Work

In this paper, we focus on the comprehensive detection of Chinese harmful memes. We present the first Chinese harmful meme dataset ToxiCN MM. It has 12k samples including not only targeted harmful memes but also those only exhibiting potential toxicity without specific targets, adapting to the Chinese online environment. In addition to binary labels, ToxiCN MM provides harmful types and modality combination categories of memes. To improve the understanding of Chinese harmful memes, we present a Multimodal Knowledge Enhancement (MKE) detector, introducing the contextual information of inline text and images. In the experimental phase, we evaluate multiple baseline models for their performance in detecting Chinese harmful memes. Our case study suggests that integrating multimodal information and comprehensive background knowledge is crucial for effective detection.

In future work, we aim to design more effective methods for Chinese harmful memes detection. Meanwhile, we notice that the accuracy of LLMs in detecting Chinese harmful memes is still limited. Considering the potential harm that these memes may cause, this task can be used to evaluate the safety of LLMs. We will employ prompt engineering and instruction fine-tuning methods to explore and enhance the detection performance of LLMs. Additionally, we will continuously evaluate state-of-the-art models to ensure the effectiveness of ToxiCN MM. We expect our dataset, benchmark, and insights will assist researchers in related fields.

7 Limitations

In this study, we focus on several most common harmful types of memes on the Chinese online environment. Due to the filtering mechanism, some harmful memes, such as those containing *fake news*, are extremely scarce on Chinese platforms. As a result, our TOXICN MM does not encompass all harmful types. The techniques we used to boost the percentage of harmful content during the dataset construction process may introduce problematic bias. In future work, we plan to broaden the scope and increase the number of meme crawls, focusing on more Chinese platforms to mitigate sampling bias. While we have implemented several measures to mitigate annotation bias, we acknowledge that our dataset may still contain mislabeled data due to the subjective understanding of annotators for Chinese harmful memes. Furthermore, our current study primarily focuses on predicting whether a given meme is harmful. We will further evaluate the ability of baselines to generate explanations for Chinese harmful memes with quantitative experiments.

8 Ethics Statement

Our study aims to facilitate the comprehensive detection of Chinese harmful memes and raise researchers' attention to non-English memes. The social psychological community has recognized the harms of the harmful types we selected in the dataset. We acknowledge the risk of malicious actors attempting to reverse-engineer memes. We strongly discourage and denounce such practices, emphasizing the necessity of human moderation to prevent them. All resources are intended solely for scientific research and are prohibited from commercial use. We believe the benefits of our proposed resources outweigh the associated risks. We strictly follow the data use agreements of each public online social platform. The opinions and findings contained in the samples of our presented dataset should not be interpreted as representing the views expressed or implied by the authors.

To mitigate the potential psychological impact on annotators evaluating harmful content, we implement the following protective measures: 1) obtain explicit consent regarding exposure to potentially abusive content, 2) limit weekly evaluations to manage exposure and ensure reasonable daily workloads, and 3) recommend discontinuing reviews if they experience distress. Additionally, we conduct regular well-being checks to monitor their mental health.

Acknowledgment

This research is supported by the Natural Science Foundation of China (No. 62376051, 62076046, 62076051), the National Language Commission Key Program (No.ZDI145-80), the Liaoning Province Applied Basic Research Program (No. 2022JH2/101300270), the Liaoning Provincial Natural Science Foundation Joint Fund Program (2023-MSBA-003), and the Fundamental Research Funds for the Central Universities (DUT24LAB123, DUT24MS003). We would like to thank all reviewers for their constructive comments.

References

Piush Aggarwal, Pranit Chawla, Mithun Das, Punyajoy Saha, Binny Mathew, Torsten Zesch, and Animesh Mukherjee. 2023. Hateproof: Are hateful meme detection systems really robust? In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 3734–3743. ACM.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

David M Bell. 1997. Innuendo. *Journal of Pragmatics*, 27(1):35–59.

Efrat Blaier, Itzik Malkiel, and Lior Wolf. 2021. Caption enriched samples for improving hateful memes detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9350–9358. Association for Computational Linguistics.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 321–332. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 67–73. ACM.

Ziyang Dong, Jinfeng Chang, and Jian Sun. 2017. A study on "dispirited culture" of online youth from the perspective of social psychology. *Youth and Adolescent Studies*, (3-7+31).

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. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 533–549. Association for Computational Linguistics.

Raul Gomez, Jaume Gibert, Lluís Gómez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 1459–1467. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 3651–3655. ACM.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Student Research Workshop, Online, November 20, 2022*, pages 32–39. Association for Computational Linguistics.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *CoRR*, abs/2308.12038.

Junhui Ji, Wei Ren, and Usman Naseem. 2023. Identifying creative harmful memes via prompt based approach. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 3868–3872. ACM.

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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Eric Koukounas and Nicole M Letch. 2001. Psychological correlates of perception of sexual intent in women. *The Journal of social psychology*, 141(4):443–456.

Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2023. Memefier: Dual-stage modality fusion for image meme classification. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, ICMR 2023, Thessaloniki, Greece, June 12-15, 2023*, pages 586–591. ACM.

Zefeng Li, Hongfei Lin, Liang Yang, Bo Xu, and Shaowu Zhang. 2022. Memeplate: A chinese multimodal dataset for humor understanding in meme templates. In *Natural Language Processing and Chinese Computing - 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24-25, 2022, Proceedings, Part I*, volume 13551 of *Lecture Notes in Computer Science*, pages 527–538. Springer.

Aijun Lin and Bo Zhang. 2019. Emojis as discourse: Symbolic consumption and sociological reflection on internet emojis. *Modern Communication (Journal of Communication University of China)*, 41(35-40).

Min Liu and Shuai Xu. 2016. An instant 'meme battle': Communication and identification in the spread of emoji packs". *Journal of News Research*, 7(339).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. Facilitating fine-grained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16235–16250. Association for Computational Linguistics.

Cunlong Miao and Maohua Xu. 2022. The representation and guidance path of youth "dispirited culture". *Journal of Socialist Theory Guide*, (123-128).

Magi Otsri. 2020. Non-sexist sexual humor as quid pro quo sexual harassment. *Sexuality & Culture*, 24(1):94–112.

Lan Peng. 2019. Emotion icon: Password, label and mask. *Journal of Xi'an Jiaotong University (Social Sciences)*, 39(104-110+153).

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2783–2796. Association for Computational Linguistics.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4439–4455. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Bjorn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *3rd Workshop on Natural Language Processing for Computer-Mediated Communication/Social Media*, pages 6–9. Ruhr-Universitat Bochum.

Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022a. DISARM: detecting the victims targeted by harmful memes. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1572–1588. Association for Computational Linguistics.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Y. Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022b. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5597–5606. ijcai.org.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

Randy Thornhill and Nancy Wilmsen Thornhill. 1983. Human rape: An evolutionary analysis. *Ethology and sociobiology*, 4(3):137–173.

Hongbo Wang, Mingda Li, Junyu Lu, Liang Yang, Hebin Xia, and Hongfei Lin. 2023. CCPC: A hierarchical chinese corpus for patronizing and condescending language detection. In *Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part II*, volume 14303 of *Lecture Notes in Computer Science*, pages 640–652. Springer.

Hongbo Wang, Junyu Lu, Yan Han, Kai Ma, Liang Yang, and Hongfei Lin. 2024a. Towards patronizing and condescending language in chinese videos: A multimodal dataset and detector. *CoRR*, abs/2409.05005.

Kaichun Wang, Junyu Lu, Bingjie Yu, Liang Yang, and Hongfei Lin. 2024b. CETA: context-enhanced and target-aware hateful meme inference method. In *Natural Language Processing and Chinese Computing - 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1-3, 2024, Proceedings, Part V*, volume 15363 of *Lecture Notes in Computer Science*, pages 95–106. Springer.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93. The Association for Computational Linguistics.

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 *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2887–2899. ACM.

Jianhua Yang. 2018. The guidance and norms of the meme culture. *People's Tribune*, (140-141).

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3181–3197. Association for Computational Linguistics.

Shengnan Zhang and Linyun Zhao. 2021. The communication mechanism and rational reflection of internet "dispirited culture". *Youth Journalist*, (40-42).

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Manning Zheng. 2016. Research on the popularity of network expression meme and its turning of discourse space. *Editorial Friend*, (42-46).

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Appendix A.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix B.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 5, Appendix C, and supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 3.5 and 4.2, and Appendix D.6.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 4.2.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix D.6.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section 5 and Appendix C.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix B.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix B.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Section 3.4.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] See Appendix B.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Appendix D.3.

A Research Background

Various harmful memes propagate on Chinese platforms (Peng, 2019; Zheng, 2016). While their creators and disseminators intend to simply express emotions or humor, and their original intent may be harmless, these memes have a significant negative impact on society when abused (Liu and Xu, 2016; Lin and Zhang, 2019). In this section, we individually explore the detrimental impacts of various harmful meme types, highlighting the importance of detecting these memes.

Targeted Harmful. Memes targeting specific individuals or groups can perpetuate hate speech, discrimination, and prejudice, fostering an environment of intolerance (Sharma et al., 2022b). They fuel online toxicity and create hostile environments. Furthermore, they have the potential to incite real-world violence or harassment and worsen social divisions.

General Offensive. Memes containing general offenses breed an aggressive culture prone to controversy and online violence (Zheng, 2016). Their influence extends beyond the individual, shaping the overall tone of the online environment. Additionally, such offensive content negatively impacts the development of correct values and healthy personalities among children (Yang, 2018).

Sexual Innuendo. While sexual innuendo content generally does not involve coercion in a sexual relationship, it can still be misconstrued due to gender and cultural differences, thereby contributing to sexualization (Thornhill and Thornhill, 1983; Koukounas and Letch, 2001). Moreover, inappropriate sexual innuendo may be considered sexual harassment (Otsri, 2020).

Dispirited Culture. Memes containing dispirited culture often evoke negative emotions and contribute to feelings of social isolation. This leads to an increase in personal depression, making it difficult for individuals to have positive interactions and relationships with others (Miao and Xu, 2022). Furthermore, the spread of these memes inadvertently undermines the value of positive thinking and promotes social anxiety (Zhang and Zhao, 2021).

B Implementation Details

B.1 Details of Data Filtering

For data filtering, we refer to the existing Chinese meme datasets (Li et al., 2022; Xu et al., 2022) and apply the following criteria:

- The meme text must contain Chinese (including code-switching); memes only containing other languages are not allowed.
- The meme text must have actual semantics; Thus, samples where the text is too brief, e.g. containing only modal particles, are removed.
- The meme must be readable. Hence, low-resolution samples that cannot be extracted inline text are excluded.
- The meme must be multimodal, meaning it should contain both the inline text and image information.

Here we present some examples of memes that were removed during the filtering process for failing to satisfy some of the above criteria, as shown in Figure B1.

(a) not in Chinese (b) limited semantics (c) blurry text (d) image-only (e) text-only

Figure B1: Examples of filtered memes and corresponding reasons. Among them, the inline text in (b) means "*ha ha ha ha ha*", which is only an onomatopoeic word without semantics.

B.2 Meme Containing Harmful Image

Based on the statistics listed in Table 3, the harmful meme where the image independently exhibits toxicity (*Harm.I*) is sparse. Here we present two samples to conduct a brief analysis, as shown in Figure B2. Both examples contain general offensive content. In Exp. (a), the *"middle finger"* is employed to convey aggression and contempt. In Exp. (b), both the inline text and image are independently harmful. The text "西内", literally translated as "*west in*" in English, serves as a homonym for *"go die"* in Japanese, while the image incorporates violent elements.

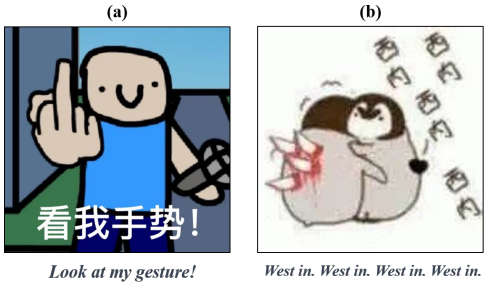

Figure B2: Examples of harmful memes where images independently exhibit toxicity (*Harm.I*).

B.3 Discussion of Annotation Consistency

After annotation, we calculate the Inter-Annotator Agreement (IAA) for each annotation hierarchy using Fleiss' Kappa values. Among them, the stage with the highest disagreement pertains to determining whether a meme is harmful, with a Kappa value of 0.62, which is comparable to other harmful meme datasets like Harm-C (0.67) and Harm-P (0.68) Blaier et al. (2021). Given the subtlety of harmful types in our dataset, this IAA is expected. These disagreements stem mainly from the humorous elements present in some samples of harmful memes, leading some annotators to consider them not toxic enough to classify them as harmful. This reflects that subjective bias still influences the results of annotation to some extent, despite implementing several measures to mitigate biases. In addition, the kappa values of discriminating harmful types and text-image combination characteristics are 0.73 and 0.86, respectively.

B.4 Statistics of Target Distribution

During the annotation phase, we label specific targets of targeted harmful memes in TOXICN MM. The target distribution is shown in Table B7.

Category	**Num**	**Ratio/%**
Gender	417	41.04
Region	176	17.32
Occupation	71	6.99
Age	66	6.50
Body	55	5.41
Race	33	3.25
Individual	29	2.85
Health	27	2.66
LGBTQ+	22	2.17
Others	120	11.81
Total	1,016	100.00

Table B7: Target distribution of targeted harmful memes in TOXICN MM.

B.5 Experimental Details

In the evaluation phase, the *harmful type identification* task is conducted as a five-classification, including *non-harmful* and four harmful types. The specific versions of each baseline are listed in Table B8. To minimize experimental error, all experiments are repeated five times.

For fine-tuned models, we acquire their original parameters from Hugging Face[2]. Weighted cross-entropy is employed to tackle category imbalances, and AdamW is selected as the optimizer. During the training phase, an early stopping mechanism is implemented to prevent overfitting. To reduce experimental error, all experiments are repeated five times with different random seeds. Details of the hyperparameter settings are presented in Table B9. All experiments are conducted using a GeForce RTX 3090 GPU.

For GPT-3.5 and GPT-4, we employ the official API provided by Openai[3] to invoke them. We respectively design the instruction templates for the two tasks, shown in Table B10 and Table B11. In addition to the definition of Chinese harmful memes, specific evaluation criteria and steps are also provided in the template that adapt to our annotation process. For example, in the evaluation criteria of the template for the task of *harmful type identification*, we emphasize that the meme containing general offense does not have specific targets, and the sexual innuendo sample does not contain sexist or sexually assault content. In future work, we plan to further optimize the design of the instruction template to improve the LLM detection performance of Chinese harmful memes. Furthermore, we will also evaluate the detection performance of the LLM in the few-shot scenario.

Model	Version
RoBERTa	`chinese-roberta-wwm-ext-base`
ResNet	`resnet-101`
ViT	`vit-base-patch16-224`
CLIP	`chinese-clip-vit-base-patch16`
GPT-3.5	`gpt-3.5-turbo`
GPT4	`gpt-4-vision-preview`

Table B8: Specific version of each baseline.

Hyperparameters	ResNet/ViT	Others
epochs	20	10
batch size	32	32
learning rate	5e-5	1e-5
padding size	-	80
dropout rate	0.1	0.1

Table B9: Hyperparameters for experiments.

C Supplementary Experiments

C.1 Performance on Diverse Modality Combination

For a supplementary analysis, we evaluate the detection performance of harmful memes with different modality combination features, as shown in Figure C1.

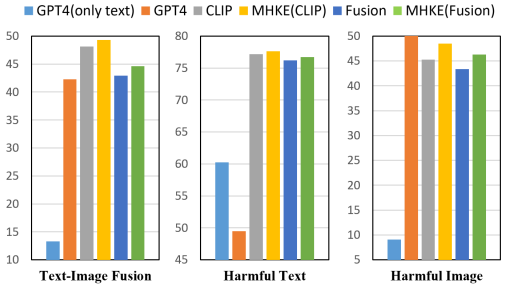

Figure C1: Accuracy of Chinese harmful memes in the test set of TOXICN MM with different modality combination features.

Compared to other combinations, memes containing harmful text are more likely to be successfully detected by the models, especially PLMs. This is because PLMs can effectively learn the unique expressions of the Chinese language during the fine-tuning stage. In contrast, for memes containing harmful images, GPT-4 demonstrates stronger performance than fine-tuned PLMs, with an average increase of 6.5%, illustrating its ability to effectively review the content of input images. Additionally, we note that GPT-4's performance on text-image fusion is comparable to that of PLMs, showcasing its capability to understand and reason with information that combines both textual and visual elements effectively. Meanwhile, compared to the model incorporating only inline text, i.e., GPT-4 (only text), GPT-4 shows a 12.7% decrease in performance for memes containing harmful text. This also supports the conclusion that benign images relatively affect the detection of GPT-4 in Chinese harmful memes. Additionally, after introducing MKE, models show improvement in detecting Chinese harmful memes with different modality combinations.

[2] `https://huggingface.co/`
[3] `https://openai.com/`

中文有害模因的定义**:**
中文有害模因是由图像和中文行内文字组成的多模态单元，具有对个人、组织、社区、社会群体或整个社会造成伤害的潜力。它们往往反映和强化了中国互联网上潜在的负面价值观和文化态度。
Definition of Chinese Harmful Meme:
Chinese harmful memes are multimodal units consisting of an image and Chinese inline text that have the potential to cause harm to an individual, an organization, a community, a social group, or society as a whole. They often reflect and reinforce underlying negative values and cultural attitudes on the Chinese Internet.

Input for Harmful Meme Detection:
给定模因 <Image>，其内联文本为 <Text> / 仅根据模因的内联文本 <Text>，将其分类为有害或无害。
评估标准:
<中文有害模因的定义> 根据定义，符合以下任何一个标准的模因将被判定为有害。
1. 模因表达了对特定个人或社会群体的厌恶、偏见或刻板印象。
2. 模因包含了讽刺或粗鲁的内容，没有特定的目标。
3. 模因暗示了性意图，以引起性冲动。
4. 模因包含了颓废和绝望的情绪，或传达了一种自我否定的态度。
评估步骤:
1. 结合图像和文本 / 仅根据文本分析模因的含义;
2. 将模因与评估标准进行比较，检查是否有任何匹配;
3. 只返回"有害"或"无害"的结论;
4. (可选) 提供做出决策的理由。
Given the meme <Image> and its inline text <Text> / Based solely on the inline text of a meme <Text>, classify it as harmful or harmless.
Evaluation Criteria:
<Definition of Chinese Harmful Meme> According to the definition, meme that meets any of the following criteria will be judged as harmful.
1. The meme expresses disgust, prejudice, or stereotypes towards specific individuals or social groups.
2. The meme encompasses sarcastic or rude content without specific targets.
3. The meme implies sexual intent to provoke sexual arousal.
4. The meme integrates decadent and desperate emotions or conveys a self-negative attitude.
Evaluation Steps:
1. Analyze the meaning of meme by combining the inline text and image / solely based on the inline text;
2. Compare the meme against the criteria to check for any matches.
3. Solely return a conclusion of "*harmful*" or "*harmless*".
4. (Optional) Provide reasons for the decision.

Table B10: Instruction templates for the task of *harmful meme detection*.

C.2 Evaluation of Chinese MLLMs

In this section, we utilize TOXICN MM to evaluate the performance of Chinese LLMs in detecting Chinese harmful memes under the zero-shot scenario. Due to the unavailability of APIs for multimodal dialogue in existing commercial Chinese LLMs (e.g., Wenxin Yiyan), we only evaluate the effect of several open-source Chinese LLMs, including VisualGLM (Du et al., 2022), Qwen-VL (Bai et al., 2023), and VisCPM (Hu et al., 2023). The results are shown in Table C1.

Model	P	R	F1
VisualGLM (ChatGLM-6B)	57.68	56.94	57.36
Qwen-VL (Qwen-7B)	59.53	58.74	58.87
VisCPM (CPMBee-10B)	54.32	54.63	54.47

Table C1: The performance of several Chinese MLLMs on the *harmful meme detection* task.

Based on the result, we note that the detection performance of these open-source Chinese LLMs is not satisfactory compared to other baselines shown in Table 5. This is because these models have weak multimodal reasoning capabilities and contain limited background knowledge required for detection.

D Licensing and Maintenance Plan

D.1 Licensing

We confirm that the dataset is licensed under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license.

中文有害模因的定义**:**
中文有害模因是由图像和中文行内文字组成的多模态单元，具有对个人、组织、社区、社会群体或整个社会造成伤害的潜力。它们往往反映和强化了中国互联网上潜在的负面价值观和文化态度。
Definition of Chinese Harmful Meme:
Chinese harmful memes are multimodal units consisting of an image and Chinese inline text that have the potential to cause harm to an individual, an organization, a community, a social group, or society as a whole. They often reflect and reinforce underlying negative values and cultural attitudes on the Chinese Internet.

Input for Harmful Type Identification
给定模因 <Image>，其内联文本为 <Text> / 仅根据模因的内联文本 <Text>，将其分类为以下其中一类。
评估标准:
<中文有害模因的定义> 根据定义，如果模因具有以下特征之一，将其分类到相应的类别：
1. 如果模因表达了对特定个人或社会群体的厌恶、偏见或刻板印象，返回"A"。
2. 如果模因包含了讽刺或粗鲁的内容，但没有特定的目标，返回"B"。
3. 模因暗示了性意图，以引起性冲动，且不包含性别歧视或性侵犯的内容，返回"C"。
4. 如果模因包含了颓废和绝望的情绪，或传达了一种自我否定的态度，返回"D"。
3. 如果模因不属于以上任何的毒性类别或无法判断，返回"E"。
评估步骤:
1. 结合图像和文本 / 仅根据文本分析模因的含义；
2. 将模因与评估标准进行比较，检查是否有任何匹配；
3. 从"A"、"B"、"C"、"D"、"E"中选择对应的类别。
Given the meme <Image> and its inline text <Text> / Based solely on the inline text of a meme <Text>, classify it into one of the categories.
Evaluation Criteria:
<Definition of Chinese Harmful Meme> According to the definition, if the meme exhibits any of the following characteristics, it will be assigned to that corresponding category.
1. If the meme expresses disgust, prejudice, or stereotypes towards specific individuals or social groups, return "A". (*Targeted Harmful*)
2. If the meme encompasses sarcastic or rude content without specific targets, return "B". (*General Offensive*)
3. If the meme implies sexual intent to provoke sexual arousal, and does not contain sexist or sexually assault content, return "C". (*sexual innuendo*)
4. If the meme integrates decadent and desperate emotions or conveys a self-negative attitude, return "D". (*Dispirited Culture*)
5. If the meme does not fall into any of the above harmful types or cannot be determined, return "E". (*Non-Harmful*)
Evaluation Steps:
1. Analyze the meaning of meme by combining the inline text and image / solely based on the inline text;
2. Compare the meme against the criteria to check for any matches.
3. Choose the corresponding category from A/B/C/D/E.

Table B11: Instruction templates for the task of *harmful type identification*.

D.2 Maintenance Plan

We will regularly update the TOXICN MM dataset by adding new samples collected from Chinese social platforms. These updates, scheduled annually, aim to maintain the dataset's currency and comprehensiveness. Additionally, we will open the *Community* section of `Huggingface` for community contributions, subjecting all submissions to rigorous review to ensure alignment with the dataset's quality standards. Moreover, based on the feedback, we will periodically enhance annotation guidelines and expand annotation types to provide a more detailed and diverse dataset. Future updates will also include enriched metadata to offer better context and support more analyses.

To address any concerns or queries related to the dataset, we will establish a dedicated support team reachable via email or through a support portal on the website. Furthermore, an issue tracking system will be implemented to document and monitor reported issues, enabling users to report bugs and suggest improvements. We encourage user feedback through a structured feedback loop, which will inform regular audits aimed at maintaining dataset integrity and quality. Periodic transparency reports will be published to keep users informed about encountered issues, resolutions, and overall dataset improvements, fostering trust and transparency within the user community.