CangjieToxi: A Chinese Offensive Language Detection Benchmark with Radical-Level Perturbations

Anonymous ACL submission

Abstract

We present **CangjieToxi**, a novel benchmark for detecting covert offensive language in Chinese social media. The dataset incorporates two real-world evasion strategies—*character splitting* and *radical substitution*—which obfuscate toxic expressions by altering the visual or structural properties of Chinese characters. These perturbations pose significant challenges for existing detection systems. To address this, we propose a **multi-stage prompting framework** that decouples character anomaly detection, semantic restoration, and toxicity classification, thereby enhancing robustness under adversarial conditions. Experiments on state-of-the-art large language models demonstrate that our method significantly outperforms baselines in both accuracy and false positive control. Our work offers a new testbed and practical mitigation strategy for building resilient toxicity detection systems. [1].

Disclaimer: *This paper describes violent and discriminatory content that may be disturbing to some readers.*

1 Introduction

In China, while social media censorship is pervasive, it is relatively less restrictive toward gender and LGBTQ+ topics compared to politically sensitive issues. These topics often resurface in "safe zones"—such as international events, public health discussions (e.g., AIDS), and the arts—where censorship is more lenient (Yu, 2024). This regulatory ambiguity allows marginalized discourse to persist, often expressed subtly through emojis, euphemisms, or references to foreign contexts (Gu and Heemsbergen, 2023). However, the same environment has become a breeding ground for gendered and LGBTQ+ hate speech, which frequently manifests in covert, lexicon-evading forms. Although censorship may not fully silence feminist or queer voices, it significantly shapes the digital landscape in which offensive language evolves and circulates.

To combat toxic speech, researchers have developed various NLP-based offensive language detection systems, especially those built upon large language models (LLMs). While effective in standard contexts, these systems consistently underperform against adversarially crafted language designed to bypass automated filters. Typical evasion strategies include homophonic substitution, emoji camouflage, and radical-level character perturbation—techniques that obscure toxicity from machines while preserving legibility for human readers (Jiang et al., 2022). For instance, the vulgar expression "操逼" can be obfuscated via character decomposition into components like "扌氵一" (Chen, 2012), or replaced by lookalike characters as in "澡称冯福" (Husain and Uzuner, 2021).

The Chinese writing system is particularly vulnerable to such manipulations due to its character-based structure and widespread reliance on lexicon-driven moderation. These evasion tactics increasingly render current moderation pipelines ineffective, allowing toxic discourse to proliferate unchecked. This reality underscores the pressing need for more robust, semantically aware detection frameworks.

In response, we introduce the **CangjieToxi** benchmark, which incorporates radical-level character decomposition and substitution perturbations to systematically evaluate model robustness. Beyond benchmarking, we propose a novel **multi-stage prompting framework** that explicitly addresses these perturbation challenges. Our approach separates the perception, restoration, and toxicity classification processes using prompt-driven LLM modules, thereby mitigating task leakage and improving both interpretability and detec-

[1] https://anonymous.4open.science/r/CangjieToxi-6D02

tion performance under adversarial conditions.

Our contributions are summarized as follows:

- We present **CangjieToxi**, a new dataset that simulates real-world evasion patterns by applying character splitting and radical substitution to offensive Chinese text.
- We propose a **multi-stage prompting-based mitigation method**, which restores perturbed characters through contextual reasoning before performing toxicity classification.
- We conduct a thorough evaluation of state-of-the-art LLMs, demonstrating that our approach significantly improves performance under both decomposition and substitution attacks.

2 Related Work

2.1 Chinese Toxic Content Detection and Dataset

Chinese toxic content detection has evolved from lexicon-based methods to advanced machine learning and large language models (LLMs). Early lexicon approaches are limited in capturing emerging or cloaked toxic expressions (Deng et al., 2022). Supervised and adversarial learning offer improved performance but remain challenged by the dynamic nature of language and the subjectivity of toxicity (Liu et al., 2023). Recent efforts in domain adaptation (Ying et al., 2024) and cross-cultural transfer (Zhou et al., 2023) have enabled the adaptation of models trained on other languages to Chinese, with promising results.

LLMs have demonstrated strong capabilities in context-aware detection. Guo et al. showed that prompt-based LLMs outperform traditional models in identifying nuanced toxic language (Guo et al., 2023), while Kumarage et al. (Kumarage et al., 2024) and Nirmal et al. (Nirmal et al., 2024) highlighted their strengths in classification and interpretability.

To support these efforts, various Chinese toxic content datasets have been developed. COLD categorizes toxicity across individual, group, and anti-bias dimensions, though with limited diversity (Deng et al., 2022). TOCP (Yang and Lin, 2020) and TOCAB (Chung and Lin, 2021) focus on profanity and abuse on Taiwan's PTT platform. SWSR targets sexism on Weibo, providing a lexicon of gender-related toxic terms (Jiang et al., 2022). ToxiCN (Lu et al., 2023), with multi-level toxicity annotations, underpins ToxiCloakCN, which addresses cloaked expressions via homophones and emoji transformations (Xiao et al., 2024).

Building on this foundation, our proposed **CangjieToxi** dataset incorporates novel perturbation strategies—such as radical-based decomposition and substitution—to challenge current models and enhance the detection of complex, cloaked toxic expressions.

2.2 LLM-based Toxicity Detection

Large language models (LLMs) have become an important direction in toxicity detection research due to their strong generalization and contextual understanding capabilities. Some applications often focused on generating or augmenting training data (Kruschwitz and Schmidhuber, 2024; Meguellati et al., 2025), other studies have increasingly investigated their direct use as classifiers. However, zero-shot or prompt-based LLM classification has shown inconsistent performance, particularly in tasks requiring nuanced social context (Meguellati et al., 2025). Moreover, LLMs may overfit to prompt phrasing or fail to generalize to implicit forms of toxicity.

Furthermore, several limitations of LLMs persist. Some studies (Zhang et al., 2024; Zhao et al., 2024) identified critical limitations of LLMs in detecting implicit hate speech, showing that they often misclassify benign statements due to oversensitivity and exhibit unreliable confidence calibration. These findings underscore the challenges of directly applying LLMs to toxicity detection and highlight the need for more robust strategies to balance sensitivity and fairness in real-world deployment.

2.3 Language Perturbation

Language perturbation techniques have been explored to examine vulnerabilities in NLP models, especially in adversarial settings. Techniques like emoji insertion (Kirk et al., 2022) and token replacement (Garg and Ramakrishnan, 2020) are commonly used to test the robustness of models against subtler forms of offensive content. In Chinese, language perturbation faces additional challenges due to the language's character-based structure, where meaning can shift dramatically with slight modifications in characters or word order. Previous work on Chinese offensive language de-

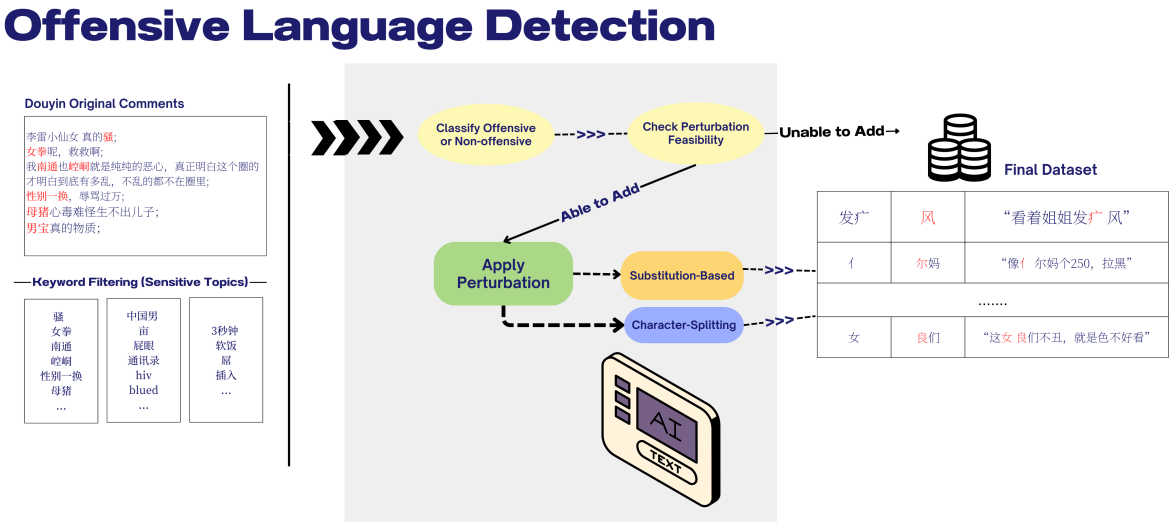

Figure 1: Offensive Langurage Detection Flowchart and Examples

tection has addressed perturbations such as word perturbation and synonym usage (Su et al., 2022), while the introduction of ToxiCloakCN demonstrates the impact of homophonic substitutions and emoji transformations on model performance (Xiao et al., 2024).

Our **CangjieToxi** dataset expands on these perturbation techniques by incorporating radical splitting and substitution of character components, adding a new layer of complexity to model testing and addressing emerging evasion tactics in Chinese offensive language detection.

3 Dataset Construction

In this section, we describe the process of constructing the dataset used for offensive language detection, including data collection, preprocessing, offensive keyword extraction, and annotation, as well as the techniques used to introduce meaningful perturbations to the dataset for training purposes. The visualization of the comprehensive process is shown in 2.

3.1 Data Source and Preprocessing

We collect comments from Douyin, a major short video platform in China. Due to the site's filtering system, posts containing offensive language are relatively rare. To address this, we focus our data collection on several sensitive topics, such as marriage, gender, fertility, LGBTQ issues, and race, which are frequently discussed online. We then compile a list of keywords for each topic and use them to gather 45484 comments that do not have replies. We exclude texts that are too short to convey meaningful content, such as those consisting only of auxiliary words or inflections. Additionally, we remove irrelevant data, such as duplicate entries and advertisements. Ultimately, 28080 comments are retained. During the data cleaning process, we removing unnecessary newlines and spaces. To protect privacy, we anonymize the data by filtering out usernames, links, emails and stickers. Since emojis may contain valuable emotional cues, we retain them for the purpose of offensive language detection.

3.2 Offensive Keywords Extraction

In order to enrich our dataset with meaningful perturbations, we applied a multi-step approach for offensive keyword extraction. First, we utilized the BERTopic model for topic modeling on our dataset, identifying offensive terms from the representative words of each topic. Additionally, we leveraged existing lexicons, such as the SexHate Lexicon from the SWSR dataset and the gender and LGBTQ+ lexicon from the ToxiCN dataset, to filter relevant offensive keywords. After filtering, we merged these external lexicons with the offensive terms we defined ourselves, creating a comprehensive keyword list, consisting of 300 offensive keywords. This lexicon was then used to screen the entire dataset for offensive content.

3.3 Human Annotation

For the annotation process, we conducted a manual review of the filtered dataset. A total of four

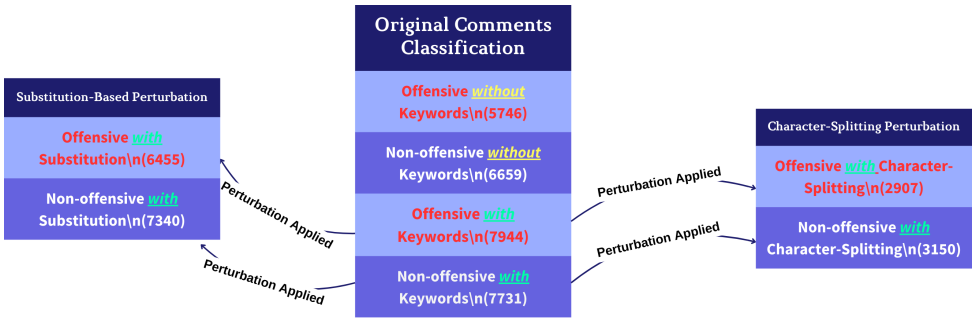

Figure 2: Offensive Langurage Detection Flowchart

native Chinese annotators with social science backgrounds were involved, ensuring gender balance in the team. To assess the reliability of the annotations, we calculated the interannotator agreement using Fleiss's Kappa, which yielded a value of 0.829, indicating a high level of agreement among the annotators. This robust agreement suggests the reliability and consistency of the offensive labels applied to the dataset.

3.4 Character-Level Perturbation

To better simulate the process of character substitution and splitting used by people to evade censorship on social media, our approach follows key principles grounded in visual recognition studies. Research has shown that substitutions or variations in character structure, as long as the distribution of information within the character remains consistent—such as maintaining the relative positions of phonetic and semantic radicals—do not significantly affect a reader's ability to recognize meaning or pronunciation (Hsiao and Cheng, 2013). This aligns with findings that visual recognition advantages in the right visual field (RVF) persist when phonetic components appear on the right and semantic components on the left, a structure commonly observed in Chinese characters (wen Hsiao, 2011). Additionally, studies on radical combinability indicate that position-specific radical combinability (SRC) is a stronger predictor of neural activation in character recognition than position-general radical combinability (GRC), suggesting that radical position matters more than sheer frequency (Liu et al., 2022). By preserving these positional relationships—especially in left-right and up-down structures—our modifications ensure that the altered characters remain easily interpretable by human readers while disrupting automated detection systems.

Our perturbation strategy differs for offensive and non-offensive text:

1. *Perturbation of offensive Text:* We only perturb words that appear in a predefined list of specific offensive keywords. This selective perturbation ensures that modifications are concentrated on words strongly associated with toxicity while avoiding unnecessary changes to unrelated words. For example, in the phrase "妈逼" (a profane expression), the character "妈" will be perturbed, whereas in "妈妈" (mother), no perturbation will occur.

2. *Perturbation of Non-offensive Text:* We perturb all individual characters that appear in the keyword list, even if they are not part of offensive words. While these perturbations are unrelated to toxicity, this design prevents the model from learning incorrect associations during training—such as mistakenly linking rare characters or structural variations with toxicity. For instance, in the word "妈妈" (mother), the character "妈" will be perturbed.

Our approach to character perturbation adheresit to three main principles:

1. *Character Structure:* We selected characters whose structure could be further split, avoiding non-split characters such as "广" (which cannot be split further). We primarily chose left-right and top-bottom structured Chinese characters, as they are the most frequently used formations in written Chinese.

2. *Position Consistency:* For both substitution and splitting, we ensured that the components retained their relative positions within the character. This structural stability minimizes disruptions in visual recognition, allowing readers to process the modified text with minimal effort.

3. *Radical Frequency:* We focused on structural components (radicals) frequently employed in character variations, ensuring that the substitutions remained consistent with real-world linguistic modifications and had minimal impact on readability.

By following these principles, our character perturbation strategy effectively mimics real-world tactics used by social media users to bypass censorship while preserving readability for human readers.

3.4.1 Character Splitting

In the Character Splitting step, we used the splitting dictionary provided by the `funnlp` library[2] to match characters in our offensive word list. The library offers multiple splitting methods for each character, and we selected the most optimal splitting method based on our principles.

The splitting rules were as follows:

1. We only split characters into two components. If a character's components exceeded two, they were placed in non-typical positions, negatively affecting recognition. For example, the character "搏" (bó) splits into '手' (hand) + '甫' (fu) + '寸' (inch), but '寸' is expected to be at the bottom of "甫," making the split unnatural.

2. When multiple splitting methods were available, we chose the method where the components' positions most closely resembled those of the original character. For instance, the character "擦" (wipe) has three splitting methods:

   - "擦" → "手" (hand) + "察" (inspect)
   - "擦" → "扌" (hand radical) + "察" (inspect)
   - "擦" → "才" (only) + "察" (inspect)

   We chose the second method because "扌" (hand radical) is most frequently seen on the left side of a character, making it the most natural and recognizable modification.[3]

3.4.2 Character Substitution

In the Character Substitution step, we relied on the library of the *Chinese Text Project* (中国哲学书电子化计划) to substitute the radical of characters from 101 offensive words, selected from a total of 300 offensive terms. These substitutions involved modifying 427 Chinese characters using different radicals.[4]

Since a single Chinese character can be substituted with multiple radicals, we followed the principle of radical frequency to determine the most suitable replacements. Specifically, we used the *Xiandai Hanyu Changyong Zibiao* (List of Frequently Used Characters in Modern Chinese) provided by the Ministry of Education [5]. Based on the individual character frequencies, we selected the most frequent substitute character with the highest frequency of occurrence as the replacement. For example, the character "猥琐" (lewd) was substituted with "偎啧" following this approach, as these substitutions closely align with commonly used radicals in modern Chinese.

This method ensures that the substitutions reflect both linguistic frequency and the intended meaning while avoiding arbitrary or non-standard replacements, helping to maintain the readability of the altered text.

4 Experiments

To evaluate the effectiveness of existing models and methods on our proposed benchmark, we employed the following experimental setup and methodologies. This systematic approach ensures a comprehensive assessment of model performance and robustness in detecting offensive language under various perturbations.

4.1 Baseline

The evaluation of three state-of-the-art models—DeepSeek-V3, GPT-4o, and Qwen-Max—revealed notable trends in their performance under character decomposition (拆字) and character substitution (换字) perturbations. On the original data, Qwen-Max achieved the highest accuracy (0.7868) and Macro F1 score (0.7858), followed by DeepSeek-V3 and GPT-4o. After applying character decomposition, all models experienced a performance decline, with DeepSeek-V3 dropping to an accuracy of 0.7165 and a Macro F1 score of 0.7150, GPT-4o dropping to an accuracy of 0.6875 and a Macro F1 score of 0.6839, and Qwen-Max dropping to an accuracy of 0.7281 and a Macro F1 score of 0.7267.

[2] https://github.com/fighting41love/funNLP

[3] https://lingua.mtsu.edu/chinese-computing/statistics/index.html

[4] https://ctext.org/dictionary.pl?if=gb

[5] https://lingua.mtsu.edu/chinese-computing/statistics/index.html

For character substitution, Qwen-Max again led with an accuracy of 0.8132 and a Macro F1 score of 0.8122, while DeepSeek-V3 and GPT-4o achieved accuracies of 0.7752 and 0.7818, respectively. The performance drop following character decomposition highlights the increased difficulty posed by this perturbation type. Notably, Qwen-Max exhibited the smallest performance degradation, suggesting stronger robustness to adversarial transformations compared to the other models. These results underscore the challenges of character-level perturbations and the varying resilience of models in handling such modifications. Detailed model performance can be seen in Table 1.

4.2 Experiment Settings

To ensure standardized and reproducible outputs from large language models (LLMs) in our experiments, we utilized the dspy framework. This framework provides a structured approach to prompt engineering and output generation, enabling consistent evaluation across different models and settings.

Prior research has demonstrated that using Chinese prompts yields marginally better performance in detecting offensive language in Chinese text compared to English prompts (Xiao et al., 2024). To align with these findings and maintain consistency, we adopted a uniform Chinese prompt across all experiments.

For all experiments involving LLMs, we set the temperature to 0.1 to minimize randomness in model outputs and ensure deterministic behavior. All other hyperparameters were kept at their default values to maintain a fair and controlled evaluation environment.

4.3 Evaluation Metric

In the field of toxic detection, the F1 score is widely regarded as the most commonly used evaluation metric, while precision is also one of the frequently employed standards in binary classification tasks. However, in previous toxic detection research, scholars have primarily focused on whether detection models can effectively identify toxic content, paying less attention to cases where normal statements are misclassified as offensive or toxic. With the growing importance of aligning large language models (LLMs) with human values, the introduction of LLMs into toxic detection tasks has made the issue of false positives more pronounced. To comprehensively evaluate the performance of different models in toxic detection, we not only utilize the F1 score but also incorporate accuracy (Acc) and false positive rate (FPR) as supplementary metrics. Among these, accuracy measures the overall performance of the model across all samples, while the false positive rate specifically evaluates the model's tendency to produce false positive classifications.

4.4 Multi-Stage Mitigation of Radical-Level Perturbations in Toxicity Detection

4.4.1 Restoring Split Characters for Toxicity Detection

To address evasion via character-splitting—where toxic characters are decomposed into component radicals to bypass detection—we propose a multi-stage restoration framework using large language models (LLMs). This method recovers disrupted semantic structures and improves toxicity classification under adversarial perturbations. The procedure involves:

Split Character Detection Given texts with character-splitting pertubations, an LLM identifies anomalous segments that may reflect character splitting. This step localizes disrupted structures indicative of adversarial intent.

Character Restoration Based on identified fragments, the LLM reconstructs the most semantically appropriate characters, guided by contextual cues. No handcrafted rules or static lexicons are used, enhancing generalizability.

Toxicity Classification The restored text is passed to an LLM classifier to assess toxicity. If restoration fails or is uncertain, the original input is used. The model applies context-sensitive reasoning to determine toxicity.

4.4.2 Recovering Substituted Characters for Robust Toxicity Detection

To mitigate radical substitution—where toxic characters are replaced with visually similar but benign variants—we adopt a complementary multi-step recovery pipeline grounded in the perturbation principles introduced in Section 3.4.2.

Perturbation Detection An LLM identifies characters whose form suggests radical-level substitution. These are flagged for possible restoration.

Candidate Retrieval Using a predefined substitution lexicon constructed in Section 3.4.2, we retrieve plausible original characters for each

Model	Data Size	GPT-4o			Qwen-Max			Deepseek-V3		
		Acc	Marco F1	FPR	Acc	Macro F1	FPR	Acc	Macro F1	FPR
Original Data	28080	0.776	0.753	0.272	0.770	0.745	0.271	0.662	0.653	0.448
Before Split	13795	0.750	0.749	0.361	0.731	0.731	0.384	0.631	0.626	0.583
After Split		0.658	0.658	0.477	0.672	0.671	0.429	0.539	0.518	0.730
Multi-Stage Mitigation Method		0.727	0.724	0.323	0.729	0.722	0.275	0.650	0.650	0.502
Before Substitution	6057	0.745	0.744	0.356	0.717	0.711	0.326	0.618	0.618	0.541
After Substitution		0.693	0.688	0.364	0.714	0.707	0.328	0.617	0.616	0.549
Multi-Stage Mitigation Method		0.766	0.716	0.086	0.745	0.694	0.103	0.681	0.680	0.426

Table 1: Model Performance in Different Conditions

anomaly, ensuring candidates maintain structural and contextual plausibility.

Contextual Replacement Selection The LLM selects the most contextually appropriate character from the candidate pool. Importantly, this step is conducted without access to the downstream toxicity objective, preserving neutrality in restoration.

Toxicity Classification The reconstructed text is then classified for toxicity using the same prompting framework. If recovery is inconclusive, the original (perturbed) text is used as fallback.

4.4.3 Unified Analysis and Design Rationale

Both mitigation pipelines follow a modular, prompt-driven architecture that decouples character recovery from toxicity classification. This design minimizes task leakage and ensures each component operates with a focused objective. By isolating restoration logic from toxicity prediction, we reduce model bias, prevent adversarial overcompensation (e.g., avoiding sensitive terms), and maintain the integrity of downstream evaluation. The shared structure across both pipelines enhances reproducibility and facilitates principled comparisons across perturbation types. The prompt we use can be found in the appendix. 2

5 Results

Table 1 summarizes the performance of three large language models—GPT-4o, Qwen-Max, and DeepSeek-V3—under both character splitting and substitution perturbation scenarios. We report Accuracy, Macro F1, and False Positive Rate (FPR) as primary evaluation metrics to assess model robustness and reliability.

5.1 Split Perturbation Results

In the "After Split" condition, all models experience a noticeable performance drop. This degradation is particularly evident in FPR, with DeepSeek-V3 rising to 0.730 and GPT-4o reaching 0.477, indicating that split-character evasion significantly hinders toxicity recognition.

Our proposed multi-stage prompting method demonstrates clear improvement across all models. For instance, Qwen-Max's FPR decreases from 0.429 to 0.275, and its Macro F1 recovers to 0.722. Similarly, GPT-4o sees a reduction in FPR from 0.477 to 0.323 while maintaining a stable accuracy.

5.2 Substitution Perturbation Results

In the substitution setting, performance degradation appears more subtle. The gap between "Before" and "After Substitution" is relatively small in terms of Accuracy and F1, but FPR remains high across all models (e.g., GPT-4o: 0.364, DeepSeek-V3: 0.549). This indicates that while models may tolerate minor visual substitutions in terms of classification, they still tend to misclassify benign text as toxic.

Our substitution recovery method notably reduces FPR while improving or maintaining classification performance. For example, GPT-4o's FPR drops dramatically to 0.086 and its Accuracy rises to 0.766. Qwen-Max also benefits, with FPR reduced to 0.103 and stable macro scores. DeepSeek-V3 shows modest improvements, validating the broad effectiveness of our approach across model families.

Overall, the multi-stage prompting strategy consistently improves robustness against both types of perturbations. It restores contextual semantics, reduces false positives, and enhances toxicity detection under adversarial conditions.

Stage	Chinese Prompt	English Prompt
Anomaly Detection	请你找出下面这段评论文本当中最为异常的一个中文汉字：	Please identify the **most anomalous Chinese character** in the following comment text:
Character Replacement	请判断文本中的异常汉字换成汉字备选中的哪一个更为合适。请分析文本语义，选择最符合中文社交媒体语境的替换方案。	Which substitute character is more appropriate for the anomalous character in the context? Analyze the semantics and select the one that best fits Chinese social media context.
Split Character Restoration	找出下面这段来自社交媒体的评论当中，可能有错误的部分（词或汉字组件）。请从汉字组成的角度，依据推断出的错误部分，纠正评论中的错误，并给出纠正后的评论。	Identify the possibly erroneous segments (words or components) in the following social media comment. Then correct the error from a character composition perspective and provide the revised comment.
Component Recomposition	请你把两个<>中间的汉字组件组合成一个新汉字（请给出一个最佳替换汉字）	Please combine the two components inside <> to form a complete Chinese character. Return the best substitution.
Toxicity Detection	请你把以下原始文本中的{异常汉字（组件）}改为{替换汉字}，然后分析评论是否是有毒的(toxic)。下面是一条来自于社交媒体的评论，在考虑到社交媒体语境的情况下，判断评论是否是"有毒的(toxic)"。	Replace the {anomalous character/component} in the original text with {replacement}, then determine whether the comment is **toxic**. Considering the social media context, classify the comment as either toxic or non-toxic.

Table 2: Multi-stage prompt design for anomaly recovery and toxicity detection. Each stage corresponds to a specific subtask in our mitigation pipeline.

6 Discussion

6.1 Lexicon and False Positive

The lexicon-based filtering approach exhibited a high false positive rate, where non-toxic content was frequently misclassified as toxic. A primary reason for this is the prevalence of comments criticizing socially undesirable behaviors (e.g., fraud, promiscuity), which, despite their harsh tone, do not constitute offensive language. This phenomenon poses a significant challenge for offensive language detection systems, as it blurs the line between legitimate criticism and actual toxicity.

To mitigate this issue, future research should prioritize the development of more advanced semantic understanding and context-aware models. Incorporating domain-specific knowledge and leveraging larger, more diverse datasets could help reduce false positives. Additionally, exploring hybrid approaches that combine lexicon-based methods with machine learning models may offer a more robust solution for distinguishing between toxic content and socially critical discourse.

6.2 Analysis of Results

The performance of Deepseek in the substitution experiment demonstrates an intriguing phenomenon: after applying the Multi-Stage Mitigation Method, its performance metrics improved compared to before the substitution. This outcome can be attributed to the nature of the substitution dataset, which is enriched with comments containing various keywords. Some of these comments, although not inherently toxic, triggered false positives in Deepseek's predictions. Following the application of the Multi-Stage Mitigation Method, the alignment of the language model with human preferences led to the rewriting of many originally toxic comments into non-toxic ones. Consequently, this method resulted in an overall improvement in the F1 score.

6.3 Future Works

Addressing offensive language that evades censorship mechanisms through techniques such as character splitting or using visually similar characters may involve two potential approaches. One approach is to employ computer vision (CV) methods to identify and associate similar characters and split characters. However, this method is costly and complicated, as the flexible structure of Chinese characters makes the problem more challenging. An alternative approach is to use "masking" techniques, which obscure key offensive terms while still allowing offensive language to be understood and recognized through contextual semantic clues—essentially enabling the system to infer meaning even when specific words are not explicitly stated (i.e., "although nothing was directly said, the intent is still understood"). The dataset we propose, which introduces perturbations only to offensive terms, is adaptable to both of these strategies.

7 Limitations

Despite the contributions made by CangjieToxi, there are several limitations in this study that should be acknowledged. First, while the dataset introduces novel perturbations such as character splitting and character substitution, it remains limited to Chinese language contexts, and the effectiveness of these evasion techniques may vary in other languages with different writing systems or character structures. Second, the perturbation methods used in this work, although effective in creating subtle forms of offensive language, are still constrained by the manual construction of these transformations, and there may be additional, unforeseen evasion tactics that were not covered. Third, the performance of state-of-the-art models on our dataset demonstrates clear limitations, but further research is needed to explore new model architectures and training methodologies that can better adapt to these types of perturbations. Finally, while we have focused on offensive language detection within social media contexts, the dataset's applicability to other domains, such as formal text or legal documents, remains to be evaluated. Future work will aim to expand these methods, explore additional types of perturbations, and assess the robustness of models across different languages and content domains.

References

Wangdao Chen. 2012. *Rhetoric introduction*. Fudan University Press. Publication date: January 1, 2008.

I Chung and Chuan-Jie Lin. 2021. Tocab: A dataset for chinese abusive language processing. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 445–452.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Yijia Gu and Luke Heemsbergen. 2023. The ambivalent governance of platformed chinese feminism under censorship: Weibo, xianzi, and her friends. *International Journal of Communication*, 17(0).

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573.

Janet H. Hsiao and Liao Cheng. 2013. The modulation of stimulus structure on visual field asymmetry effects: The case of chinese character recognition. *The Quarterly Journal of Experimental Psychology*, 66(9):1739–1755. PMID: 23391072.

Fatemah Husain and Ozlem Uzuner. 2021. A survey of offensive language detection for the arabic language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(1).

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Hannah Kirk, Bertie Vidgen, Paul Rottger, Tristan Thrush, and Scott Hale. 2022. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368, Seattle, United States. Association for Computational Linguistics.

Udo Kruschwitz and Maximilian Schmidhuber. 2024. LLM-based synthetic datasets: Applications and limitations in toxicity detection. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 37–51, Torino, Italia. ELRA and ICCL.

Tharindu Kumarage, Amrita Bhattacharjee, and Joshua Garland. 2024. Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. *Preprint*, arXiv:2403.08035.

Hanyu Liu, Chengyuan Cai, and Yanjun Qi. 2023. Expanding scope: Adapting English adversarial attacks to Chinese. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 276–286, Toronto, Canada. Association for Computational Linguistics.

Xiaodong Liu, David Wisniewski, Luc Vermeylen, Ana F. Palenciano, Wenjie Liu, and Marc Brysbaert. 2022. The representations of chinese characters: Evidence from sublexical components. *Journal of Neuroscience*, 42(1):135–144.

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), pages 16235–16250, Toronto, Canada. Association for Computational Linguistics.

Elyas Meguellati, Assaad Zeghina, Shazia Sadiq, and Gianluca Demartini. 2025. Llm-based semantic augmentation for harmful content detection. *Preprint*, arXiv:2504.15548.

Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. Towards interpretable hate speech detection using large language model-extracted rationales. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 223–233, Mexico City, Mexico. Association for Computational Linguistics.

Hui Su, Weiwei Shi, Xiaoyu Shen, Zhou Xiao, Tuo Ji, Jiarui Fang, and Jie Zhou. 2022. RoCBert: Robust Chinese bert with multimodal contrastive pretraining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–931, Dublin, Ireland. Association for Computational Linguistics.

Janet Hui wen Hsiao. 2011. Visual field differences in visual word recognition can emerge purely from perceptual learning: Evidence from modeling chinese character pronunciation. *Brain and Language*, 119(2):89–98. Neurocognitive Processing of the Chinese Language.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. ToxiCloakCN: Evaluating robustness of offensive language detection in Chinese with cloaking perturbations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6012–6025, Miami, Florida, USA. Association for Computational Linguistics.

Hsu Yang and Chuan-Jie Lin. 2020. TOCP: A dataset for Chinese profanity processing. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 6–12, Marseille, France. European Language Resources Association (ELRA).

Hao Ying, Qiongrong Ou, Chengjun Fan, Lin Mei, Shuyu Zhang, and Xu Xu. 2024. Domain adaptation fornbsp;chinese offensive language detection. In *Natural Language Processing and Chinese Computing: 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1–3, 2024, Proceedings, Part IV*, page 146–158, Berlin, Heidelberg. Springer-Verlag.

Jinyang Yu. 2024. *Shifting shadows: media attention and censorship of gay people in China (1949-2023)*. Ph.D. thesis, University of British Columbia.

Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of LLMs in implicit hate speech detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12073–12086, Bangkok, Thailand. Association for Computational Linguistics.

Yibo Zhao, Jiapeng Zhu, Can Xu, and Xiang Li. 2024. Enhancing llm-based hatred and toxicity detection with meta-toxic knowledge graph. *Preprint*, arXiv:2412.15268.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023. Cross-cultural transfer learning for Chinese offensive language detection. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 8–15, Dubrovnik, Croatia. Association for Computational Linguistics.

A APPENDIX