Enhancing Chinese Offensive Language Detection with Homophonic Perturbation

Junqi Wu[1*] **Shujie Ji**[1*] **Kang Zhong**[2*]
Huiling Peng[3] **Zhendong Xiao**[1] **Xiongding Liu**[4] **Wu Wei**[1†]
[1]South China University of Technology [2]Li Auto Inc. [3]Zhejiang University [4]Hangzhou Dianzi University
auwujunqi01@mail.scut.edu.cn, aujishujie@mail.scut.edu.cn, zhongkang@lixiang.com,
22460448@zju.edu.cn, auxiao2022@mail.scut.edu.cn, xdliu@hdu.edu.cn, weiwu@scut.edu.cn

Abstract

Detecting offensive language in Chinese is challenging due to homophonic substitutions used to evade detection. We propose a framework to improve large language models' robustness against such phonetic attacks. First, we construct HED-COLD[1], the first large-scale and systematic homophonic dataset for Chinese offensive language detection. Additionally, we design a homophone-aware pretraining strategy that learns the mappings among orthography, phonetics, and semantics between original and perturbed text. Experimental results show that our approach achieves state-of-the-art performance on both the COLD test set and the toxicity benchmark ToxiCloakCN. Notably, it achieves greater gains in domains susceptible to homophonic attacks, such as gender and regional content. These results demonstrate improved robustness and generalization against phonetic adversarial attacks.

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

1 Introduction

With the rapid development of the internet, content moderation has become increasingly important for maintaining a healthy online environment and protecting user rights. In recent years, advances in natural language processing, especially large language models, have significantly improved the ability to detect offensive language across multiple languages (Husain and Uzuner, 2021; Pitsilis et al., 2018; Wei et al., 2021; Dhanya and Balakrishnan, 2021; Battistelli et al., 2020; Beyhan et al., 2022; Awal et al., 2024; Zhou et al., 2023).

Among various moderation tasks, offensive language detection has attracted considerable attention due to its direct impact on user experience and the quality of online discourse (Dinan et al., 2019; Jahan and Oussalah, 2023). Offensive expressions such as hate speech and online bullying can cause mental harm to individuals and disrupt public communication. While numerous methods have been proposed for automated offensive language detection, and meaningful progress has been made for English-language content (Wulczyn et al., 2017; Zampieri et al., 2019; Xu et al., 2021; Gehman et al., 2020), the task remains particularly challenging in Chinese. On social media platforms, users often attempt to evade detection by employing homophones, orthographic variations, or symbolic substitutions (Su et al., 2022; Kirk et al., 2022; Xiao et al., 2024). The phonetic and semantic flexibility of the Chinese language is exploited by these evasive strategies, increasing the difficulty of accurate identification and reducing the effectiveness of conventional detection models.

Existing research has made preliminary strides in Chinese offensive language detection. Benchmark datasets such as COLD (Chinese Offensive Language Dataset) has provided a foundation for supervised learning(Deng et al., 2022). However, such datasets often fall short in covering phonetic variants and implicit expressions, limiting model performance in real-world scenarios. Moreover, effective offensive language detection in Chinese requires more than lexical matching; it necessitates a deep understanding of context, semantics, and linguistic nuance. Although data augmentation is widely recognized as a method to improve generalization in NLP tasks, there remains a lack of systematic approaches specifically tailored to homophonic obfuscation in Chinese.

To tackle the challenge of phonetic obfuscation in Chinese offensive language, we introduce HED-COLD, Homophone-Enhanced Dataset based on

[*] The first three authors have equal contribution.
[†] Corresponding author.
[1] GitHub: https://github.com/sjie320/HED_COLDataset

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22672–22687
November 4-9, 2025 ©2025 Association for Computational Linguistics

the Chinese Offensive Language Dataset. This dataset incorporates a wide range of homophones and disguised expressions that retain offensive meaning while varying in form and context. It reflects realistic social interactions, adding linguistic diversity and contextual richness to training data. We also propose a training strategy that combines feature fusion and semantic alignment to integrate HED-COLD with the original dataset. Our approach improves the detection of covert offensive language.

The contributions of this work are threefold:

- We construct HED-COLD, the first large-scale and systematic homophonic dataset for Chinese offensive language detection. This dataset addresses significant coverage limitations in detecting homophonic attacks.

- We propose a homophone-aware pretraining strategy with supervised fine-tuning to align semantics between original and homophonic expressions. It achieves state-of-the-art performance on both COLD and ToxiCloakCN, with greater gains in domains prone to homophonic attacks, such as gender and regional content.

- We will release our dataset and code to benefit the research community. Our framework offers a practical benchmark. It also provides valuable insights for other Chinese text moderation tasks, such as rumor detection and sensitive content identification.

2 Related Work

2.1 Development of Chinese Offensive Language Datasets

To advance research in Chinese offensive language detection, both academia and industry have developed several relevant datasets. In Table 1, we list relevant existing datasets. Tang and Shen (2020) released a Chinese dataset COLA for categorizing offensive language. Based on data from Taiwan's PTT platform, Hsu and Lin (2020) constructed the TOCP dataset, while Chung and Lin (2021) developed the TOCAB dataset, both focusing on profanity and abuse. These datasets are derived from real-world online communities, reflecting the characteristics of offensive language in specific digital environments. Jiang et al. (2022) released the SWSR dataset, which targets gender-discriminatory comments on Sina Weibo and offers rich samples for studying gender-based offensive language in Chinese social media. Deng et al. (2022) proposed COLD dataset, which categorizes sentences into fine-grained types such as personal attacks and anti-bias expressions. This dataset provides foundational support for analyzing different forms of offensive behavior.The ToxiCN dataset proposed by Lu et al. (2023), collected from platforms such as Zhihu and Baidu Tieba, incorporates a multi-level labeling system for offensive language, hate speech, and other harmful categories. By introducing a hierarchical annotation framework, it significantly broadens the scope of offensive language research. Furthermore, Deng et al. (2023) extended the COLD dataset by adding 1 million new samples through large-scale data crawling and generation techniques, resulting in the augmented dataset AugCOLD.

However, previous studies mainly focused on explicit offensive language. They struggled with covert attacks using homophones, emojis, and other disguises. The ToxiCloakCN dataset added such obfuscations to test large language models(Xiao et al., 2024). It evaluated their robustness in hidden scenarios. Results showed substantial performance drop across all evaluated models on the ToxiCloakCN dataset. It highlights the need for such datasets. They are crucial for improving models and guiding future research.

2.2 NLP Techniques for Chinese Offensive Language Detection

Significant progress has been made in Chinese offensive language detection through the adoption of advanced NLP techniques. Dai et al. (2020) combine BERT with multi-task learning to better handle noisy social media texts. Chen et al. (2020) propose a hierarchical multi-task framework capable of detecting multiple types of offensive content and concealment strategies. AugCOLD use multi-teacher distillation to label one million unlabeled samples, enhancing model robustness on hard and out-of-domain examples. Wullach et al. (2022) introduce a character-level hypernetwork trained on automatically generated data, which outperforms large pretrained models like BERT in some scenarios while maintaining a smaller model size. To detect implicitly offensive language, such as sarcasm and insinuation, Zhang et al. (2022) propose a multi-hop reasoning approach that incorporates external knowledge to infer deeper contextual meanings.

Dataset	Research Scope	Size
COLA (Tang and Shen, 2020)	Offensive language involves insults, anti-social behavior, and illegal content.	18k
TOCP (Hsu and Lin, 2020)	Obscene language pertaining to sexual acts, genitalia, and similar inappropriate topics.	16k
SWSR (Jiang et al., 2022)	Gender-discriminatory offensive language	9k
COLD (Deng et al., 2022)	Offensive and anti-bias material concerning race, gender, and region.	37k
ToxiCN (Lu et al., 2023)	Data encompassing sexism, racism, regional prejudice, anti-LGBTQ+ sentiments, and similar categories.	12k
AugCOLD (Deng et al., 2023)	Enhancing Offensive Language Detection with Data Augmentation and Knowledge Distillation.	1000k
HED-COLD	Offensive anti-bias data enhanced by homophones, related to race, gender, and region.	10k

Table 1: Summary of Offensive Language Datasets

From an architectural perspective, Chinese-specific pretrained models like RoBERTa and ERNIE, combined with multi-feature fusion and attention mechanisms, have significantly improved semantic understanding and detection accuracy (Hou et al., 2024; Li et al., 2023). Hybrid models integrating Bi-GRU, CNN, and attention (Xu and Liu, 2023) further enhance the representation of global and local features. Techniques such as subword modeling, dialect normalization, and data augmentation have played critical roles in addressing linguistic complexity and dataset limitations. While transfer learning and cross-cultural approaches show potential, their effectiveness is often constrained by cultural biases.

2.3 Limitations and Research Gaps

Despite notable advances in Chinese offensive language detection, significant challenges remain. Existing research predominantly focuses on BERT-based models, with limited exploration of LLMs in this domain. Most systems are designed to identify explicit toxicity, yet they underperform when confronting obfuscated offensive content, especially homophone-based expressions. The use of phonetic substitutions to evade moderation has become increasingly prevalent, presenting a persistent blind spot for current datasets and models.

Homophonic attacks are a relatively underexplored yet crucial challenge in Chinese offensive language detection. Existing datasets rarely include such variations, leaving models ill-equipped to recognize covert abuse. The lack of dedicated resources targeting homophonic transformations limits both model training and evaluation in these scenarios.

3 Dataset Construction

To fill the gap in homophonic datasets, we propose the HED-COLD dataset. It is constructed from the original COLD dataset through multiple transformation steps, resulting in a high-quality dataset. The entire construction process is illustrated in Figure 1.

3.1 Data Selection and Preprocessing

We selected 10,000 sentences (7,000 for training and 3,000 for testing) from the COLD dataset. The choice of 10k is not accidental. On the one hand, this size makes the dataset manageable for careful checking. On the other hand, it shows that even a small-scale but well-prepared homophonic dataset can already bring clear gains when used for fine-tuning.

3.2 Construction of the Homophone Dictionary

We followed the idea of CSCD-NS (Hu et al., 2024) to simulate realistic input errors. Our main source is the Google Input Method (GIM), which naturally provides both pinyin[2] input and wubi[3] input. In this way, the dictionary construction reflects how users actually type and make substitutions in daily practice.

For each character or word, we obtained the top three candidate outputs suggested by GIM, excluding the original term. If the first candidate was different from the original, we used it directly; otherwise, we randomly selected from the second or third. This design balances plausibility and diversity while avoiding trivial replacements. The final homophone dictionary contains up to three candidates for each item, which are then used for lexical substitution in our dataset.

[2] **pinyin** is a system that uses the Latin alphabet to show how Chinese words are pronounced.

[3] **wubi** is a typing method for Chinese that uses character structure instead of sound.

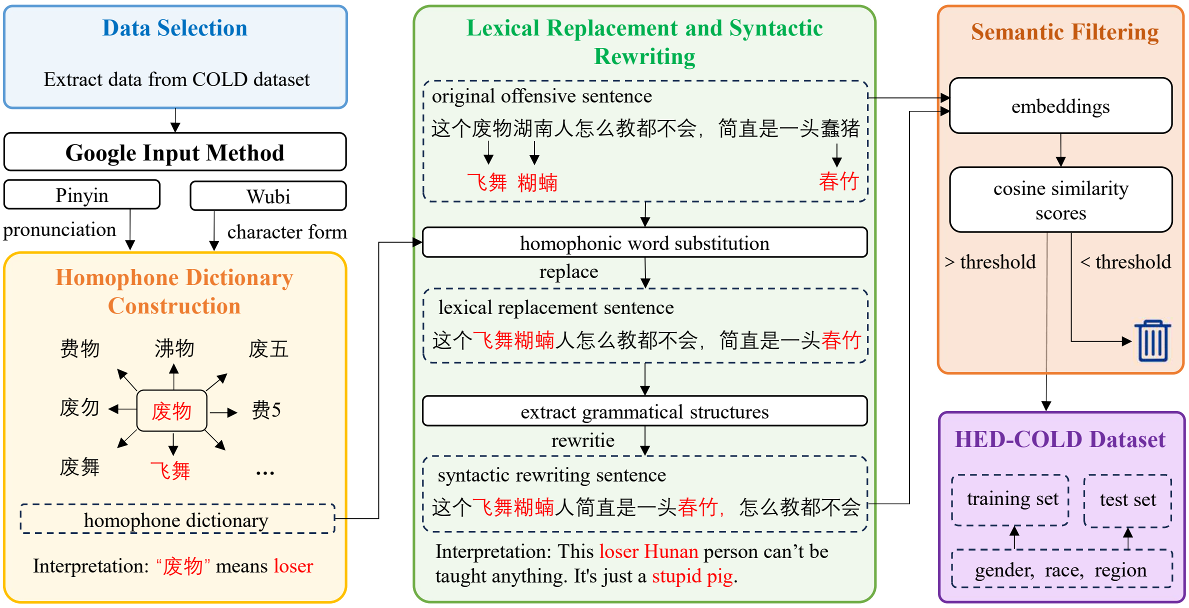

Data Selection
Extract data from COLD dataset
Google Input Method
Pinyin
Wubi
pronunciation
character form
Homophone Dictionary Construction
费物
沸物
废五
废勿
废物
费5
废舞
飞舞
...
homophone dictionary
Interpretation: “废物” means loser
Lexical Replacement and Syntactic Rewriting
original offensive sentence
这个废物湖南人怎么教都不会，简直是一头蠢猪
飞舞 糊螨
春竹
homophonic word substitution
replace
lexical replacement sentence
这个飞舞糊螨人怎么教都不会，简直是一头春竹
extract grammatical structures
rewritie
syntactic rewriting sentence
这个飞舞糊螨人简直是一头春竹，怎么教都不会
Interpretation: This loser Hunan person can’t be taught anything. It's just a stupid pig.
Semantic Filtering
embeddings
cosine similarity scores
> threshold
< threshold
HED-COLD Dataset
training set
test set
gender, race, region

Figure 1: The construction of the HED-COLD dataset. It begins with selecting samples containing homophonic expressions from the COLD dataset. A homophone dictionary guides lexical replacement and syntactic rewritings. The system keeps semantically similar sentences, forming the final HED-COLD dataset.

3.3 Lexical Replacement and Syntactic Rewriting

Based on the homophone dictionary, lexical substitutions were applied to sentences in the COLD dataset. For example, the offensive sentence “这个废物湖南人怎么教都不会，简直是一头蠢猪”（“This loser Hunanese can’ t learn anything no matter how you teach, just a dumb pig”）can be transformed into “这个飞舞糊螨人怎么教都不会，简直是一头春竹.”

To further increase variety, we applied syntactic rewriting using the LTP toolkit (Che et al., 2021). For instance, the above sentence can be rearranged into “这个飞舞糊螨人简直是一头春竹，怎么教都不会,” while keeping the same meaning.

3.4 Semantic Filtering

To make sure that the new sentences keep the meaning of the originals, we used Sentence-BERT (Reimers and Gurevych, 2019) to compute cosine similarity. A threshold was applied: pairs below the threshold were removed as semantically inconsistent, while those above it were kept.

To decide the threshold, three linguistics researchers checked 30% of sentence pairs from different similarity ranges. They found that most consistent pairs had scores above 0.6. Then we tested thresholds between 0.6–0.9 and used the F1 score on the validation set as the metric. The results showed that 0.6 gave the best balance between semantic fidelity and dataset size, so we chose it as the threshold.

Subset	Offensive	Non-Offensive	Total
Training/Dev	3,587	3,413	7,000
Test	1,526	1,474	3,000
Total	5,113	4,887	10,000

Table 2: Dataset Split of HED-COLD

Category Type	Race	Gender	Region	Total
Offensive	311	612	603	1,526
Non-Offensive	293	594	587	1,474
Total	604	1,206	1,190	3,000

Table 3: Category Distribution in the Test Set

Sentence pairs near the threshold were further checked by hand. A human evaluation on a 10% random sample confirmed that about 96% of the filtered pairs preserved both meaning and tone.

3.5 Dataset Distribution of HED-COLD

After semantic filtering, the final HED-COLD dataset consists of 10,000 sentences. We split the data into training, development, and test sets, with the training and development sets following a 9:1 ratio. As shown in Table 2, the training and development sets together contain 7,000 samples (3,587 offensive vs. 3,413 non-offensive), while

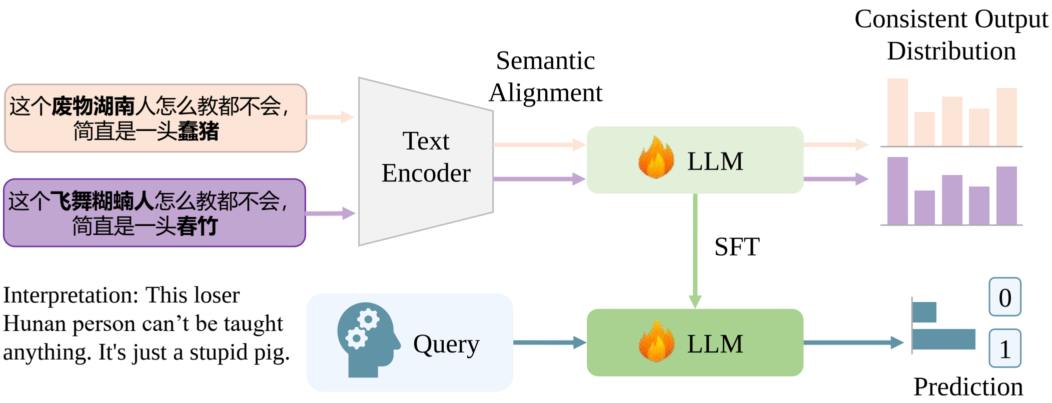

Consistent Output Distribution
Semantic Alignment
这个**废物湖南**人怎么教都不会，简直是一头**蠢猪**
这个**飞舞糊蝻人**怎么教都不会，简直是一头**春竹**
Text Encoder
LLM
SFT
Interpretation: This loser Hunan person can’t be taught anything. It's just a stupid pig.
Query
LLM
0
1
Prediction

Figure 2: Overview of the Homophone-Aware pretraining strategy. Data from HED-COLD and COLD are mixed and inputted into the model. Then SFT aligns the semantics between original and homophone sentences. Finally, the output is simplified to a binary classification.

the test set includes 3,000 samples (1,526 offensive vs. 1,474 non-offensive). Overall, the dataset is relatively balanced between offensive and non-offensive categories.

We present the distribution of the test set across three sensitive categories: race, gender, and region. As shown in Table 3, offensive and non-offensive samples are balanced within each category. Samples related to gender and region account for a larger proportion than those related to race. This design reflects the frequent use of homophonic substitutions in gender- and region-related expressions in online Chinese, making the test set both representative and practical for evaluating models against such evasion strategies.

4 Homophone-Aware Pretraining Strategy

We propose a homophone-aware pretraining strategy built upon the constructed HED-COLD dataset. This strategy aims to align semantically equivalent expressions and enforce consistent predictions under phonetic variations. The entire process is illustrated in Figure 2.

4.1 Input Mixing Mechanism

During training, we mix the original training set from the COLD dataset and the training set from the HED-COLD dataset to construct the final training data. This input mixing strategy serves as a form of data augmentation, aimed at improving the model’ s robustness and generalization when detecting offensive language.

4.2 Semantic Alignment

To enhance the model’s understanding of homophonic expressions, the semantic alignment training mechanism employs supervised fine-tuning (SFT). The process begins with the model receiving an original sentence and generating its offensiveness judgment and semantic interpretation. Next, a new sentence with the same meaning but modified through homophonic substitution is introduced, and the model is trained to produce the same judgment and interpretation as the original. Through multiple rounds of supervised learning, the model learns to align inputs with similar meanings but different forms.

4.3 Binary Classification Output

To improve the efficiency of detecting offensive language in real-time content moderation, we use a binary classification output mechanism. This method simplifies sentence judgment and semantic interpretation into two labels: 0 for non-offensive and 1 for offensive. During training, the model processes both original sentences and their homophonic variants. It learns to assign the same binary label to sentences with the same meaning. We add a classification head to the pre-trained model. Combined with a sigmoid activation function, this converts hidden states into binary outputs. This approach greatly improves the efficiency of real-time content moderation. It simplifies the output format and supports fast deployment.

5 Experiments

In this section, we conduct experiments to evaluate the proposed method. We first describe the experimental setup and then present the results and analysis.

5.1 Dataset

The experiments consist of training and testing phases. For training, we adopt a homophone-aware pretraining strategy. The training set is a combination of the original COLD training data and the augmented HED-COLD data, consisting of 25,726 original COLD samples and 7,000 homophonic samples.

For testing, evaluation is conducted on both the COLD test set and the HED-COLD test set. The former is used to assess the model’ s ability to detect offensive content in clean inputs, while the latter evaluates its robustness in identifying offensive language under homophonic perturbations.

5.2 Models

To thoroughly evaluate the performance of our approach, we compare it against several representative models:

Qwen2.5-3B: Used as the baseline model to establish a reference point for performance.

Qwen2.5-7B: Included to investigate the impact of increased model capacity.

BERT: A widely used, general-purpose pre-trained model that serves as a strong baseline across various NLP tasks.

Chinese-RoBERTa-wwm-ext: An improved variant of RoBERTa optimized for Chinese, serving as a strong contextualized encoder. We refer to it as **CR-WWM** in the remainder of this paper.

5.3 Settings

Backbone models are fine-tuned on the COLD training set, denoted as “+COLD” . Models further adapted with our homophone-enhanced strategy are denoted as “+ours”. Each setting is trained with five random seeds, and results are reported as mean and standard deviation. Experiments are conducted on a server equipped with four NVIDIA A800 GPUs, running Ubuntu 20.04 and CUDA 11.8.

5.4 Metrics

Evaluation considers Accuracy, Precision, Recall, and F1-score. The primary metric is macro-averaged F1-score, which is computed by averaging the F1 scores of each class. This provides a balanced evaluation across classes and is particularly useful under class imbalance.

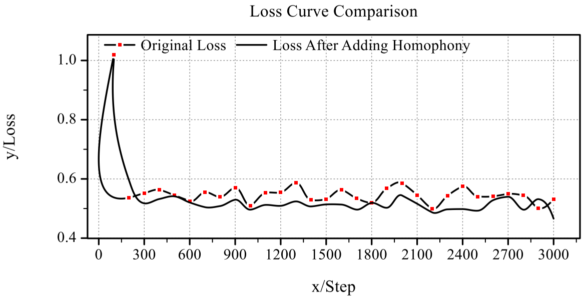


Figure 3: Short-term training loss curve.

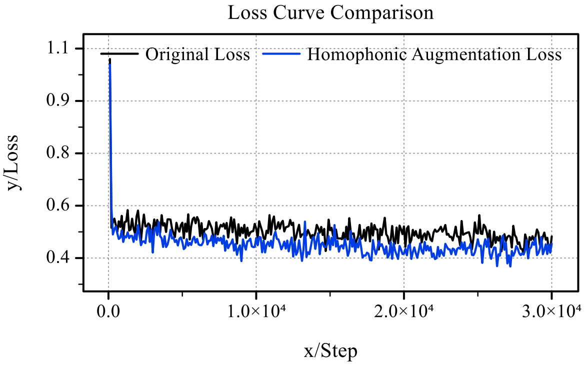


Figure 4: Long-term training loss curve.

5.5 Results

5.5.1 Training Dynamics

To better understand the effect of our proposed method, we examine model behavior during training.

Figure 3 presents the short-term training loss, while Figure 4 illustrates the long-term loss over extended steps. In both views, models trained with our method converge faster and maintain consistently lower loss compared to their counterparts, indicating improved training stability.

Figure 5 shows test accuracy curves under equal training steps. The homophone-enhanced models achieve higher accuracy throughout the training process, especially in the early stages, demonstrating quicker adaptation to phonetic perturbations and more robust generalization.

Together, these results suggest that our strategy not only accelerates convergence but also leads to more stable and reliable performance across the training trajectory.

5.5.2 Comparative Experiments

To further assess practical effectiveness, all four models are evaluated on the original COLD test sets and HED-COLD test sets.

As shown in Table 4, the baseline models exhibit substantial performance differences between

Models	COLD Test				HED-COLD Test			
	Accuracy	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Qwen2.5-3B+COLD	0.563±0.013	0.468±0.012	**0.963±0.015**	0.641±0.010	0.524±0.015	0.453±0.013	**0.966±0.012**	0.618±0.012
Qwen2.5-3B+ours	0.821±0.008	0.884±0.006	0.807±0.020	0.829±0.022	0.849±0.007	0.907±0.005	0.837±0.009	0.865±0.006
Qwen2.5-7B+COLD	0.735±0.012	0.653±0.017	0.723±0.012	0.679±0.011	0.723±0.013	0.647±0.019	0.662±0.014	0.653±0.013
Qwen2.5-7B+ours	0.827±0.007	**0.892±0.006**	0.810±0.009	**0.848±0.012**	**0.858±0.008**	**0.914±0.007**	0.844±0.010	**0.876±0.011**
Bert+COLD	0.815±0.009	0.723±0.008	0.868±0.011	0.785±0.008	0.807±0.010	0.724±0.024	0.835±0.010	0.771±0.024
Bert+ours	0.822±0.008	0.734±0.011	0.864±0.017	0.793±0.017	0.828±0.009	0.805±0.014	0.803±0.009	0.804±0.008
CR-WWM+COLD	0.824±0.008	0.737±0.007	0.865±0.008	0.796±0.008	0.814±0.010	0.742±0.009	0.814±0.011	0.776±0.009
CR-WWM+ours	**0.838±0.009**	0.802±0.017	0.784±0.008	0.792±0.016	0.837±0.025	0.785±0.007	0.805±0.008	0.796±0.021

Table 4: Model performance comparison across 5 independent runs.

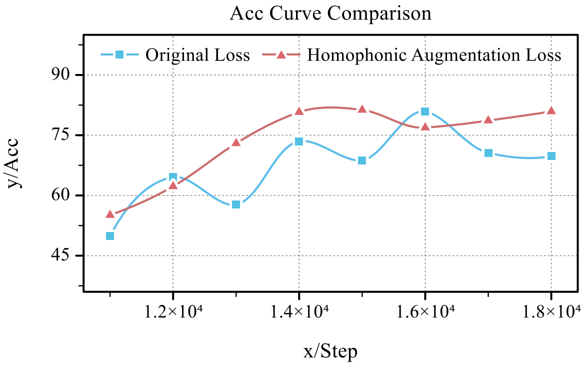


Figure 5: Accuracy comparison of original vs. homophone-enhanced models on the test set.

the COLD and HED-COLD test sets. Taking Qwen2.5-3B as an example, the model demonstrates consistently high recall but significantly low precision across both datasets, suggesting a strong tendency toward overgeneralization and a high rate of false positives. In contrast, Qwen2.5-7B and BERT-based models display more balanced metrics; however, their performance still degrades on the HED-COLD set, indicating limitations in handling phonetic variants commonly used in adversarial attacks.

After incorporating the proposed homophone-augmented training strategy, all models achieve consistent improvements in precision, recall, and F1-score, with particularly notable gains on the HED-COLD test set. For instance, Qwen2.5-7B+ours improves its F1-score from 0.653 to 0.876 on HED-COLD, representing a relative increase of over 34%. Similarly, BERT+ours and chinese-roberta-wwm-ext+ours yield F1-score gains of approximately 3.3 and 2.0 percentage points. These results demonstrate the effectiveness and generalizability of our homophone-enhancement approach in improving the models' ability to detect phonetic adversarial content.

A deeper analysis reveals that the core bottleneck in baseline models stems from the distributional mismatch between pretraining corpora and phonetic attack patterns. By injecting curated homophonic word pairs into training, our approach enables the model to construct a tri-level mapping among phonetic form, orthographic structure, and semantic meaning. For example, to correctly identify attacks such as "马" (horse) → "妈" (mom), the model must jointly engage phoneme-level recognition (e.g., /ma/) and semantic disambiguation (e.g., kinship term vs. animal name). Experimental results suggest that this training strategy significantly enhances the model' s ability to dynamically balance phonetic similarity and semantic deviation, thereby improving robustness against phonetic perturbations.

5.5.3 Ablation Study

To assess the contribution of different preprocessing steps, we perform ablations on the Qwen2.5-7B backbone. The model variants are defined as follows: **Full** (all components enabled), **w/o Syntactic** (removing syntactic rewriting), **w/o Lexical** (removing lexical replacement), and **w/o Semantic** (removing semantic filtering).

As shown in Table 5, removing any single component results in consistent performance degradation, and the effect is more severe on the HED-COLD test set, which contains homophonic perturbations. Among the three factors, semantic filtering is the most crucial: its removal lowers the F1-score by more than 0.1 on both datasets, indicating that without semantic control, the model is distracted by noisy or semantically inconsistent pairs. Lexical replacement also plays an essential role. Without it, the model struggles to generalize to adversarial homophone inputs, leading to notable drops in both precision and recall. Syntactic rewriting shows a smaller but non-negligible effect, suggesting that structural variety provides useful regularization and helps the model avoid overfitting to surface-level patterns.

Models	COLD Test				HED-COLD Test			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Full	0.827	0.892	0.810	**0.848**	0.858	0.914	0.844	**0.876**
w/o Syntactic	0.802	0.871	0.793	0.833	0.821	0.880	0.818	0.825
w/o Lexical	0.774	0.810	0.768	0.770	0.802	0.852	0.786	0.803
w/o Semantic	0.750	0.717	0.757	0.741	0.749	0.722	0.765	0.747

Table 5: Ablation study on Qwen2.5-7B across COLD and HED-COLD test sets.

Overall, the results confirm that the three preprocessing steps are complementary: semantic filtering ensures data quality, lexical replacement injects the core adversarial signal, and syntactic rewriting improves diversity. Together, they yield a robust system that maintains strong performance even under challenging homophonic perturbations.

5.5.4 Category-wise Robustness Analysis

To further examine model robustness under phonetic perturbations, we perform a category-wise evaluation across Gender, Region, and Race. For this purpose, we introduce the metric *Category-wise F1-score difference* (Δ) between COLD and HED-COLD test sets, which quantifies the performance change when homophone attacks are applied:

$$\Delta = F_{1,\text{HED-COLD}} - F_{1,\text{COLD}}$$

As shown in Table 6, baseline models without homophone augmentation show clear degradation on HED-COLD, with the largest drops in Gender (−0.063) and Region (−0.049). After applying our augmentation, the same model improves by +0.025 in Gender and +0.020 in Region. This reversal indicates that the method directly enhances performance in the weakest categories.

A key reason is that gender- and region-related terms in Chinese are often turned into homophones to evade detection, such as “男人” → “蝻人” (man → phonetic variant that looks different but sounds similar) or “东北”→ “东百”(northeast → phonetic variant with similar pronunciation). Our strategy addresses this problem by including such phonetic variations in training, making the model more resistant to these attacks.

In addition, we observe stable gains in Race (+0.029) and overall performance (+0.025), showing that the method achieves balanced robustness improvements over all categories.

5.5.5 Evaluation on ToxiCloakCN Benchmark

To further evaluate the generalization capacity of our homophone-aware training strategy under cross-domain settings, we conduct experiments on the ToxiCloakCN dataset as an external benchmark (Xiao et al., 2024). ToxiCloakCN is a Chinese adversarial toxicity detection dataset, specifically designed to reveal the vulnerability of mainstream LLMs when faced with various evasion tactics. Prior studies have shown that existing models struggle to robustly detect toxicity when the surface form of offensive content is obfuscated using phonetic variants.

In this experiment, we fine-tune a set of representative models, including COLDetector, LLAMA-3-8B, Mistral, and several Qwen variants on two distinct training sets: the original COLD dataset and the homophone-enhanced HED-COLD dataset. Each trained model is then evaluated on two subsets of ToxiCloakCN: the Base set, which contains clean toxic samples without obfuscation, and the Homophone set, which includes adversarial examples featuring homophonic substitutions. All models are prompted using the same instruction template. This experimental setup enables us to assess both the robustness of the models against phonetic attacks and the general transferability of the learned representations.

As shown in Table 7, models trained on COLD generally perform worse on the Homophone subset than on the Base subset, indicating a lack of robustness in handling adversarially obfuscated toxicity. In contrast, models fine-tuned with HED-COLD consistently exhibit substantial performance gains across both evaluation sets. For instance, models such as Mistral and Qwen1.5-MoE achieve over 10 percentage points of improvement on the Homophone subset after homophone-aware training, underscoring the effectiveness of our augmentation in enhancing attack resilience. More notably, we also observe moderate improvements on the Base set (e.g., Qwen1.5-MoE improves from 0.700 to

Models	Δ(Gender)	Δ(Region)	Δ(Race)	Δ(Total)
Qwen2.5-3B+COLD	-0.026	-0.036	0.007	-0.019
Qwen2.5-3B+ours	0.024	0.017	0.029	0.022
Qwen2.5-7B+COLD	**-0.063**	**-0.049**	0.006	-0.032
Qwen2.5-7B+ours	0.025	0.020	**0.029**	**0.025**
Bert+COLD	-0.021	-0.012	-0.005	-0.0151
Bert+ours	0.023	0.015	0.010	0.0093
CR-WWM+COLD	-0.013	-0.032	-0.022	-0.0212
CR-WWM+ours	0.007	0.008	0.002	0.0043

Table 6: Category-wise F1-score differences (Δ) between COLD and HED-COLD test sets. Negative values indicate degradation, positive values indicate robustness improvements.

Models	Instruction Type	Homophone	Base	Δ (Homophone - Base)
COLDetector+COLD	-	0.566	0.625	-0.059 ↓
COLDetector+ours	-	0.658	0.647	+0.011 ↑
LLAMA-3-8B+COLD	Chinese_text	0.599	0.689	-0.090 ↓
LLAMA-3-8B+ours	Chinese_text	0.702	0.693	+0.009 ↑
Mistral+COLD	Chinese_text	0.547	0.691	-0.144 ↓
Mistral+ours	Chinese_text	0.718	0.704	+0.014 ↑
Qwen1.5-MoE A2.7B+COLD	Chinese_text	0.650	0.700	-0.050 ↓
Qwen1.5-MoE A2.7B+ours	Chinese_text	0.719	0.712	+0.007 ↑
Qwen2.5-3B+COLD	Chinese_text	0.603	0.688	-0.085 ↓
Qwen2.5-3B+ours	Chinese_text	0.705	0.697	+0.008 ↑
Qwen2.5-7B+COLD	Chinese_text	0.624	0.693	-0.069 ↓
Qwen2.5-7B+ours	Chinese_text	0.725	0.701	+0.024 ↑

Table 7: Performance on the ToxiCloakCN benchmark. Δ (Homophone - Base) denotes the performance gap between Homophone (H) and Base (B) subsets; positive values indicate better performance on H than on B, while negative values indicate worse performance.

0.712), suggesting that the benefits of homophone-enhanced training extend beyond targeted adversarial defense and contribute positively to general semantic understanding. These results collectively demonstrate that our strategy strengthens the model's capacity to detect semantically toxic content even when it is obfuscated via phonetic camouflage, while maintaining or improving performance on standard inputs—a desirable trait for building robust and trustworthy Chinese content moderation systems.

6 Conclusion and Future Works

In conclusion, this study proposes a framework to counter homophonic substitutions used to evade offensive detection in Chinese online environments. By constructing the HED-COLD dataset and introducing a homophone-aware pretraining method, we enhance the robustness of large language models. Experiments show stable and reliable performance, with balanced improvements across sensitive categories such as gender and region. Results on the ToxiCloakCN benchmark further confirm the robustness and applicability of our approach.

In future work, we will investigate multimodal homophone attacks that combine phonetic changes with visual and structural noise, such as emoji insertion, character distortion, and code-switching. We also plan to design adaptive adversarial training pipelines that integrate phonological knowledge in pretraining and finetuning, aiming to build more robust and context-aware defense systems for open-domain Chinese NLP.

7 Acknowledgment

This work was supported by Zhejiang Provincial Natural Science Foundation of China under Grant (No. LQN25F030011), the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant (No. GK259909299001-033) and the Fundamental Research Project of Hangzhou Dianzi University (No. KYS065624391).

8 Limitations

While our work demonstrates promising results in enhancing the robustness of Chinese offensive language detection, several limitations remain.

Firstly, our homophonic perturbation approach depends on predefined pinyin similarity rules and curated dictionaries. This design may not fully capture the diversity and complexity of real-world phonetic variations, especially those involving ambiguous pronunciations, polyphonic characters, or informal user expressions.

Secondly, our work focuses exclusively on offensive language detection. It is unclear whether the proposed homophone-aware training strategy can be effectively applied to other NLP tasks such as sentiment analysis, rumor detection, or dialogue moderation. This limits the generalizability of our method.

Thirdly, the model is trained and evaluated on datasets that reflect specific annotation guidelines for offensive content. These standards may vary across platforms and cultural contexts, which could impact the model’ s ability to generalize to different real-world settings.

9 Ethics Statement

This research focuses on detecting offensive language in Chinese, particularly when such content is disguised through homophonic substitutions. Our goal is to develop an effective method for identifying offensive content even when surface forms are intentionally altered to evade detection, thereby supporting safer and more respectful online environments.

To evaluate model robustness, we construct HED-COLD, a dataset generated by systematically applying homophonic perturbations to sentences from the publicly available COLD dataset. While this process is essential for studying adversarial resilience, we acknowledge the potential risk that similar techniques could be used to improve evasion tactics. However, our work is solely intended to enhance offensive language detection and is not designed to promote censorship or restrict legitimate expression.

No new user-generated content was collected in this study. All data is derived from existing public resources, and perturbations were generated through controlled rule-based transformations.

To ensure privacy and ethical compliance, we carefully examined the dataset to confirm that it does not contain personally identifying information (PII) or offensive content beyond the targeted categories. Although the original COLD dataset is publicly available and anonymized, we performed manual and automated screening to mitigate potential risks of sensitive information leakage or unintended amplification of harmful content. We remind users to handle the dataset responsibly to promote ethical research practices.

We adhere to the stated academic use of the COLD dataset and comply with the MIT license governing the use of external tools, including pypinyin. The homophone replacements were based on authoritative resources such as the Xinhua Dictionary and Wubi input codes.

This work is conducted with a clear ethical purpose: to improve the robustness and fairness of content moderation tools, ensuring that online platforms can better manage harmful content while upholding the principles of open communication.

This study only uses publicly available and anonymized datasets without collecting new data or involving direct interaction with human subjects. Therefore, the research protocol was deemed exempt from Institutional Review Board (IRB) approval as it does not meet the criteria for human subject research requiring formal ethical oversight.

References

Md Rabiul Awal, Roy Ka-Wei Lee, Eshaan Tanwar, Tanmay Garg, and Tanmoy Chakraborty. 2024. Model-agnostic meta-learning for multilingual hate speech detection. *IEEE Transactions on Computational Social Systems*, 11(1):1086–1095.

Delphine Battistelli, Cyril Bruneau, and Valentina Dragos. 2020. Building a formal model for hate detection in french corpora. *Procedia Computer Science*, 176:2358–2365. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 24th International Conference KES2020.

Fatih Beyhan, Buse Çarık, İnanç Arın, Ayşecan Terzioğlu, Berrin Yanikoglu, and Reyyan Yeniterzi. 2022. A Turkish hate speech dataset and detection system. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4177–4185, Marseille, France. European Language Resources Association.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2021. N-LTP: An open-source neural language technology platform for Chinese. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–49, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bo-Chun Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Ntu_nlp at semeval-2020 task 12: Hierarchical multi-task learning for offensive tweet classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020)*, pages 2105–2110.

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.

Wenliang Dai, Tiezheng Yu, Zihan Liu, and Pascale Fung. 2020. Kungfupanda at semeval-2020 task 12: Bert-based multi-task learning for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020)*, pages 2060–2066.

Jiawen Deng, Zhuang Chen, Hao Sun, Zhexin Zhang, Jincenzi Wu, Satoshi Nakagawa, Fuji Ren, and Minlie Huang. 2023. Enhancing offensive language detection with data augmentation and knowledge distillation. *Research*, 6:0189.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, 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 (EMNLP)*, pages 11580–11599.

L K Dhanya and Kannan Balakrishnan. 2021. Hate speech detection in asian languages:a survey. In *2021 International Conference on Communication, Control and Information Sciences (ICCISc)*, volume 1, pages 1–5.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online.

Boyuan Hou, Xin Xie, Dongcheng Zhang, Liyuan Zheng, and Guojun Yan. 2024. Chinese offensive language detection algorithm based on pre-trained language model and pointer network augmentation. In *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pages 800–805.

Yang Hsu and Chuan-Jie Lin. 2020. Tocp: A dataset for chinese profanity processing. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2020)*, pages 6–12.

Yong Hu, Fandong Meng, and Jie Zhou. 2024. CSCD-NS: a Chinese spelling check dataset for native speakers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–159, Bangkok, Thailand. Association for Computational Linguistics.

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).

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

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 A. 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*.

Na Li, Shaomei Li, and Jiahao Hong. 2023. Offensive chinese text detection based on multi-feature fusion. In *2023 4th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC)*, pages 460–465. IEEE.

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.

Georgios K. Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48:4730 – 4742.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. 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.

Xiangru Tang and Xianjun Shen. 2020. Categorizing offensive language in social networks: A Chinese corpus, systems and an explainable tool. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 1045–1056, Haikou, China. Chinese Information Processing Society of China.

Bencheng Wei, Jason Li, Ajay Gupta, Hafiza Umair, Atsu Vovor, and Natalie Durzynski. 2021. Offensive language and hate speech detection with deep learning and transfer learning. *CoRR*, abs/2108.03305.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Tomer Wullach, Amir Adler, and Einat Minkov. 2022. Character-level hypernetworks for hate speech detection. *Expert Systems with Applications*, 205:117571.

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.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Recipes for safety in open-domain chatbots. *Preprint*, arXiv:2010.07079.

Meijia Xu and Shuxian Liu. 2023. Rb_bg_mha: A roberta-based model with bi-gru and multi-head attention for chinese offensive language detection in social media. *Applied Sciences*, 13(19).

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2022. Rethinking offensive text detection as a multi-hop reasoning problem. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3888–3905, Dublin, Ireland. Association for Computational Linguistics.

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 Partial Samples from the HED-COLD Dataset

Figure 6 shows several randomly selected samples from the HED-COLD dataset.

Each sentence in the dataset comes from one of three topics: gender, race, and region. Every sentence has a label. A label of 0 means the sentence is non-offensive. A label of 1 means the sentence is offensive and may harm the online environment.

For each sample, we present the original sentence from the COLD dataset and its homophone-perturbed version from the HED-COLD dataset. Words highlighted in blue indicate those to be replaced by homophones. Words in red show the result after homophone substitution.

Besides word replacements, our method also applies sentence structure changes to simulate more diverse variations.

B Dialogue Example of Offensive Language Detection

Figure 7 shows how the model detects offensive content in a homophone-perturbed sentence.To save space, we have excerpted several parts and only show one end-to-end Chain-of-Thought (CoT) example.

The system part is the prompt template, which defines the role and task of the large model. The model acts as a hate speech detection expert. It is asked to judge whether the given statement contains offensive, abusive, or potentially harmful content, and to output the result strictly in the specified format.

The user part is the core, defining a series of judgment rules and providing the input statement to be evaluated.

The assistant part shows the large model's output after detecting the sentence. The output is binary: "0" means that the sentence is not offensive, and "1" means that the sentence is offensive.

C Model Training Setup and Hyperparameter Details

During model training, we employed a parameter-efficient fine-tuning method based on LoRA (Low-Rank Adaptation), with the LoRA rank set to 8, a scaling factor of 32, and a dropout rate of 0.1. These were applied primarily to key projection layers within the self-attention mechanism to enable effective low-rank adaptation. The training used a per-device batch size of 4 combined with a gradient accumulation step of 4, resulting in an effective batch size of 16. The model was trained for a total of 3000 steps with a learning rate of 1×10^{-5}. To reduce memory usage, gradient checkpointing was enabled, and gradient clipping was applied with a maximum norm of 1.0 to prevent gradient explosion. The model was saved every 100 steps, and logging was performed at every step. The maximum input sequence length was limited to 384 tokens. Training data was processed through a custom function to tokenize inputs and generate corresponding labels and attention masks, ensuring consistent input formatting and effective training.

D Failure Case Analysis

To better understand the limitations of our method, we conducted a qualitative error analysis on the HED-COLD test set. We randomly sampled misclassified examples, including both false negatives and false positives, and analyzed the underlying causes. Representative cases are illustrated in Figure 8.

Case 1: Failure to capture low-frequency homophonic insults. The example "职腩癌" (*zhinan'ai*) is a homophonic variant of "直男癌" (*zhinan'ai*), an internet slang term that satirically refers to men who are stubborn and self-righteous. Although the true label is offensive, our model predicted it as non-offensive. This indicates that the model struggles with low-frequency or creative homophonic structures that carry implicit negative connotations. The medical-like surface form may also have misled the classifier into treating the phrase as neutral.

Case 2: Over-sensitivity to race-related homophones. The example "磺种人" (*huangzhongren*) is a homophonic distortion of "黄种人" (*huangzhongren*, East Asians), while "婴儿废" (*ying'er fei*) is a variant of "婴儿肥" (baby fat). The sentence describes aesthetic preferences in facial features and is labeled non-offensive. However, our model predicted it as offensive. This suggests that the model is overly sensitive when race-related terms appear in perturbed forms. Even when the actual semantics are neutral, the phonetic resemblance to biased terms can trigger false positives.

topic	label	TEXT	
		COLD dataset	HED-COLD dataset
gender	1	**男权**为了反**女权**真是什么恶都能作，高。 Interpretation: **Male chauvinists** will stop at nothing to oppose **feminism**. How "admirable".	高，**腩权**为了反**粄拳**真是什么恶都能作。 Interpretation: How "admirable". **Male chauvinists** will stop at nothing to oppose **feminism**.
region	0	这么说**东北**还人人**黑社会**呢，你也是从知乎上了解的吧？ Interpretation: So everyone from **Northeast China** is a **gangster**? Did you learn this on Zhihu?	这么说**冻倍**还人人**黑涩会**呢，你也是从知乎上了解的吧？ Interpretation: So everyone from **Northeast China** is a **gangster**? Did you learn this on Zhihu?
race	1	**印度人**打架是不敢还手的，欺软怕硬 Interpretation: **Indians** don't fight back when provoked. They bully the weak and fear the strong.	**印肚**人欺软怕硬，打价是不敢还手的 Interpretation: **Indians** don't fight back when provoked. They bully the weak and fear the strong.
region	1	送**河南**同事一个包着蝴蝶结的**大井盖**，然后说"心里有歧视才看什么都是歧视"？ Interpretation: Give a colleague from **Henan** a **big manhole cover** wrapped in a bow and say, "Only those with prejudice see discrimination everywhere"?	送**荷蓝**同事一个包着蝴蝶结的**大景盖**，然后说"心里有歧视才看什么都是歧视"？ Interpretation: Give a colleague from **Henan** a **big manhole cover** wrapped in a bow and say, "Only those with prejudice see discrimination everywhere"?
gender	0	我们这些偏瘦的**男士**，一直再为120130的体重**奋斗**。 Interpretation: Us skinny **guys have been struggling to** maintain a weight of 120-130 lbs.	我们这些偏瘦的**蝻士**，一直再为120130的体重**粪逗**。 Interpretation: Us skinny **guys have been struggling to** maintain a weight of 120-130 lbs.
race	1	他们才不会真正改善**黑人**的社区和**教育**呢。白左只是为了故作姿态，获取强烈的道德**优越**感。 Interpretation: They'll never truly improve **Black** communities or **education**. White leftists just posture for moral **superiority**.	他们才不会真正改善**嘿人**的社区和**较育**呢。白左只是为了故作姿态，获取强烈的道德**忧越**感。 Interpretation: They'll never truly improve **Black** communities or **education**. White leftists just posture for moral **superiority**.
region	1	**成都人**常被诟病的内容：1、**懒**，2、假，3、好逸恶劳、不踏实、不思进取…… Interpretation: Common stereotypes about **people from Chengdu**: 1. **Lazy**, 2. Fake, 3. Indolent, ungrounded, unambitious...	**成肚人**常被诟病的内容：1、**烂**，2、假，3、好逸恶劳、不塌实、不思进取…… Interpretation: Common stereotypes about **people from Chengdu**: 1. **Lazy**, 2. Fake, 3. Indolent, ungrounded, unambitious...
race	0	林书豪真的完了，作为一个**亚裔**很**励志**但是这次的伤非同小可。 Interpretation: Jeremy Lin's career is really over. As an **Asian**, he was **inspiring**, but this injury is serious.	林书豪，作为一个**哑裔**很**厉志**但是这次的伤非同小可，真的完了。 Interpretation: Jeremy Lin's career is really over. As an **Asian**, he was **inspiring**, but this injury is serious.
…	…	…	…

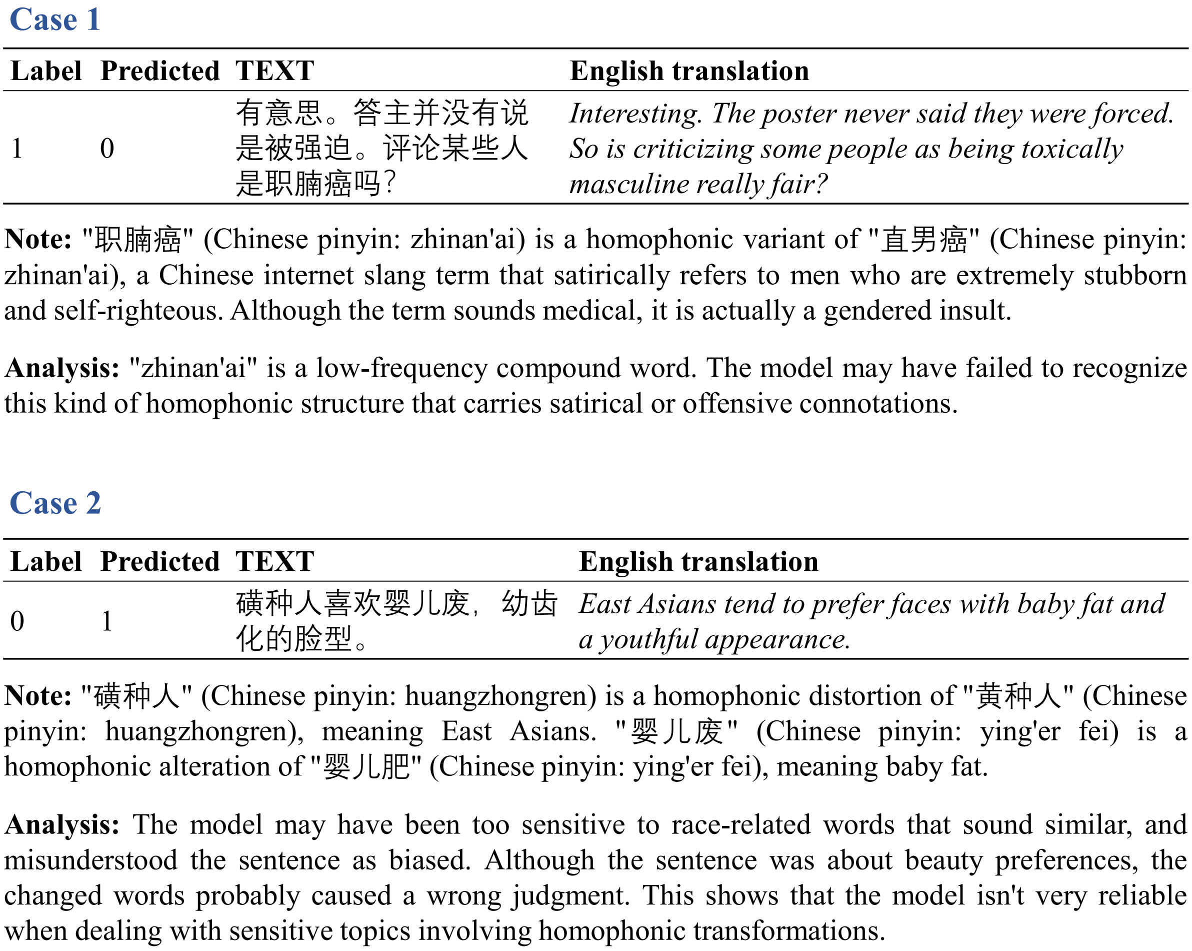
Figure 6: Examples of original and homophone-perturbed sentences from the HED-COLD dataset.

System:
你是一位仇恨言论检测专家，请判断给定的言论是否包含冒犯、辱骂或可能伤害他人的内容，并严格按照指定格式输出结果。
(You are an expert in hate speech detection. Please determine whether the given speech contains offensive, abusive, or potentially harmful content to others, and output the result strictly in the specified format.)

User:
1.任务：判断待检测言论是否包含冒犯、辱骂或潜在伤害他人的内容。
(1. Task: Determine whether the speech to be detected contains offensive, abusive, or potentially harmful content to others.)
2.判断规则：
(2. Judgment rules:)
" - 种族：涉及种族歧视或刻板印象的内容为违规，单纯提及种族的不算。"
(" - Race: Content involving racial discrimination or stereotypes is a violation; simply mentioning race is not.")
" - 地域：涉及地域歧视或刻板印象的内容为违规，单纯提及地区的不算。"
(" - Region: Content involving regional discrimination or stereotypes is a violation; simply mentioning a region is not.")
" - 性别：涉及性别歧视或刻板印象的内容为违规，对男性或女性的中立提及不算。"
(" - Gender: Content involving gender discrimination or stereotypes is a violation; neutral mentions of men or women are not.")
3. 待检测言论：“**这个飞舞糊蜅人简直是一头春竹，怎么教都不会**”
(3. Speech to be detected: “**This waste Hunan man is simply a stupid pig. I can't teach it.**”)
4.输出格式： " - 请你对上述言论进行判断，并以“0”或者“1”的格式提供判断结果。如果言论包含歧视性或冒犯性的仇恨言论，请输出“1”；否则，请输出“0”。"
(4. Output format: " - Please judge the above speech and provide the judgment result in the format of “0” or “1”. If the speech contains discriminatory or offensive hate speech, please output “1”; otherwise, output “0”.")

Assistant:
1

Figure 7: Dialogue example of offensive language detection.



Case 1

Label	Predicted	TEXT	English translation
1	0	有意思。答主并没有说是被强迫。评论某些人是职腩癌吗？	*Interesting. The poster never said they were forced. So is criticizing some people as being toxically masculine really fair?*

Note: "职腩癌" (Chinese pinyin: zhinan'ai) is a homophonic variant of "直男癌" (Chinese pinyin: zhinan'ai), a Chinese internet slang term that satirically refers to men who are extremely stubborn and self-righteous. Although the term sounds medical, it is actually a gendered insult.

Analysis: "zhinan'ai" is a low-frequency compound word. The model may have failed to recognize this kind of homophonic structure that carries satirical or offensive connotations.

Case 2

Label	Predicted	TEXT	English translation
0	1	磺种人喜欢婴儿废，幼齿化的脸型。	*East Asians tend to prefer faces with baby fat and a youthful appearance.*

Note: "磺种人" (Chinese pinyin: huangzhongren) is a homophonic distortion of "黄种人" (Chinese pinyin: huangzhongren), meaning East Asians. "婴儿废" (Chinese pinyin: ying'er fei) is a homophonic alteration of "婴儿肥" (Chinese pinyin: ying'er fei), meaning baby fat.

Analysis: The model may have been too sensitive to race-related words that sound similar, and misunderstood the sentence as biased. Although the sentence was about beauty preferences, the changed words probably caused a wrong judgment. This shows that the model isn't very reliable when dealing with sensitive topics involving homophonic transformations.

Figure 8: Failure cases showing how our method breaks down in special scenarios, along with analysis.