Enhancing Chinese Offensive Language Detection with Homophonic Perturbation

Anonymous ACL submission

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, a homophone-enhanced dataset based on the Chinese Offensive Language Dataset. Additionally, we propose a homophone-aware pretraining strategy that aligns semantics and fuses features to learn robust mappings 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 especially prone to homophonic attacks, such as gender and regional content. These results demonstrate improved robustness and generalization against phonetic adversarial attacks.

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., 2023; 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 (Noever, 2018; 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 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, a Chinese homophone offensive language dataset. 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.

Table 1: Summary of Offensive Language Datasets

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

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.

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 samples from the COLD dataset, including 7,000 from the training set and 3,000 from the test set. This dataset contains Chinese sentences annotated as either offensive or non-offensive. These samples were chosen due to their high potential for phonetic manipulation, as they frequently include words or phrases that can be substituted with homophones commonly used in offensive language.

3.2 Construction of the Homophone Dictionary

To accommodate the linguistic characteristics of Chinese, we constructed an initial phonetic-shape mapping table based on the Xinhua Dictionary of Chinese Homophones[1]. To ensure high-quality substitutions, we applied a two-tier filtering strategy: (1) phonetic similarity measured by pinyin edit distance[2], and (2) orthographic similarity assessed by prefix matching in Wubi input codes[3]. For each Chinese character, the top three most plausible homophonic candidates were identified. A manual review phase followed, during which semantically ambiguous candidates were excluded. The result is a refined, high-quality homophone dictionary used for substitution tasks.

3.3 Lexical Replacement and Syntactic Rewriting

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

[1]**Xinhua Dictionary** is a widely used Chinese language dictionary, often used in schools and education. It provides standard pronunciations and character meanings.

[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.

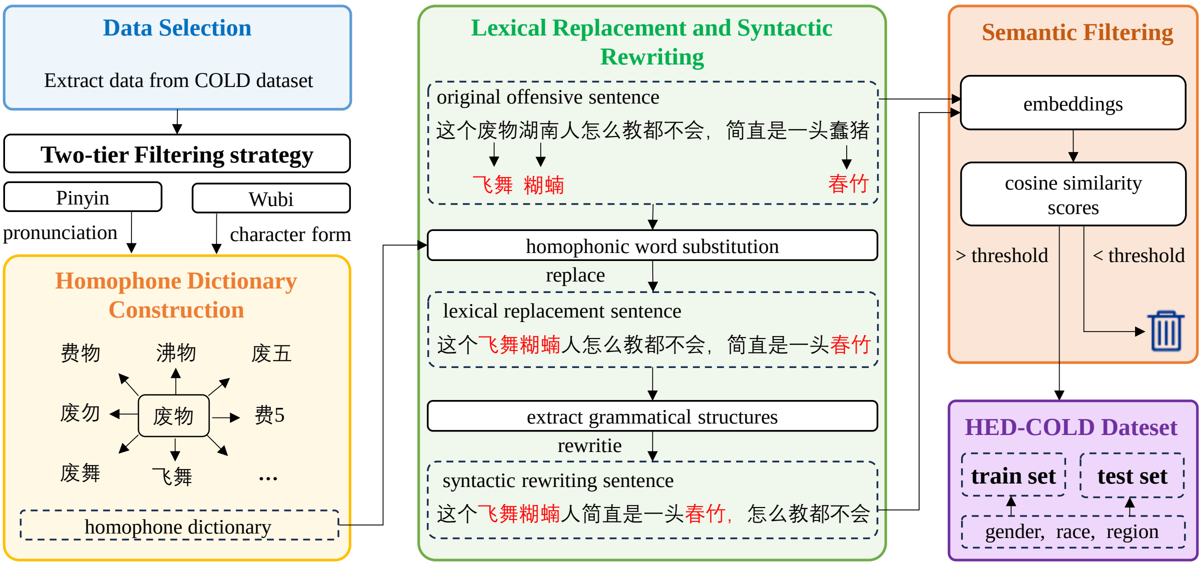

Figure 1: The construction of the HED-COLD dataset. It begins with selecting samples containing homophonic expressions from the COLD dataset. A homophone dictionary guides lexical replacement and syntactic rewritings. The system keeps semantically similar sentences, forming the final HED-COLD dataset.

蝻人怎么教都不会，简直是一头春竹”, where words are replaced with similar-sounding but obfuscated characters.

To further increase linguistic variety, we applied syntactic paraphrasing techniques to the homophonically perturbed sentences. Specifically, we used the LTP toolkit developed by HIT (Che et al., 2020) to extract grammatical structures and applied a set of syntactic transformation rules to generate alternate formulations. For instance, the sentence above could be rearranged into “这个飞舞糊蝻人简直是一头春竹，怎么教都不会” while preserving its original semantics.

3.4 Semantic Filtering

To ensure semantic consistency between the original and transformed sentences, we employed pretrained language models to generate sentence embeddings for both. We then calculated cosine similarity scores between each original–transformed pair. A similarity threshold was applied to retain only those homophonic sentences whose semantic content closely matched that of the original. The threshold value was empirically determined using a small set of manually labeled semantically consistent sentence pairs, with fine-tuning conducted to identify the optimal cutoff point.

The final HED-COLD dataset, derived from filtered sentences, comprises 10,000 samples focusing on gender, region, and race. It contains a training set with 7,000 samples and a test set with 3,000 samples.

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.

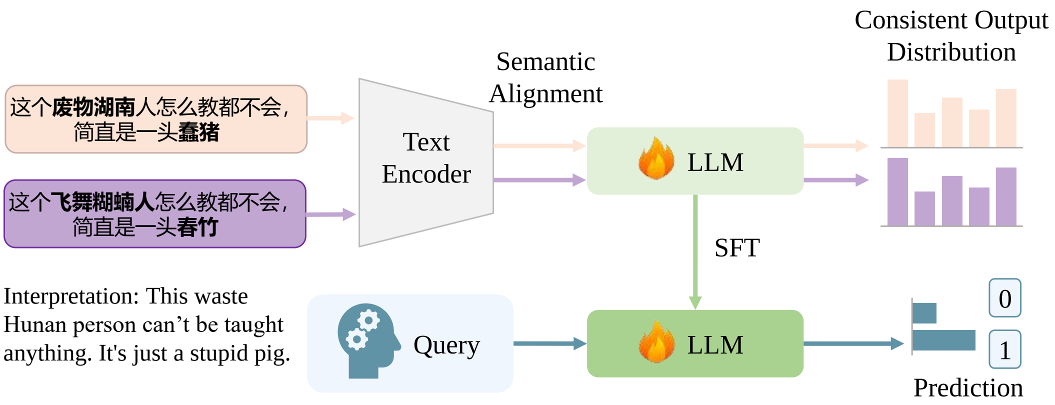

Figure 2: Overview of the Homophone-Aware pretraining strategy. Data from HED-COLD and COLD are mixed and inputted into the model. Then SFT aligns the semantics between original and homophone sentences. Finally, the output is simplified to a binary classification.

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

5.1 Experimental Setup

5.1.1 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.

5.1.2 Contrast Systems

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.

5.1.3 Settings

On the basis of these backbone models, we further apply our proposed homophone-aware fine-tuning strategy. The resulting models are denoted as **XXX+ours**, where **XXX** refers to the corresponding base model.

Experiments are conducted on a server with four NVIDIA A800 GPUs, running Ubuntu 20.04 and CUDA 11.8.

5.1.4 Metrics

Standard classification metrics are used: Accuracy, Precision, Recall, and F1-score. Among them, F1-score is the primary metric to comprehensively evaluate model robustness under homophone interference.

5.2 Experimental Results

5.2.1 Training Dynamics Observation

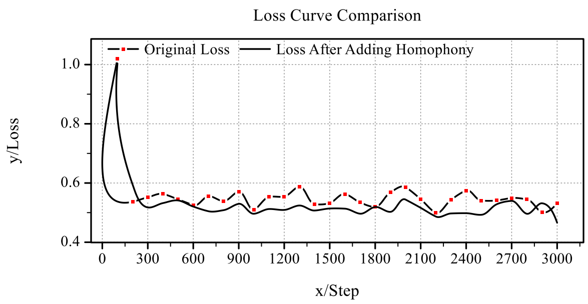

Figure 3: Short-term training loss curve.

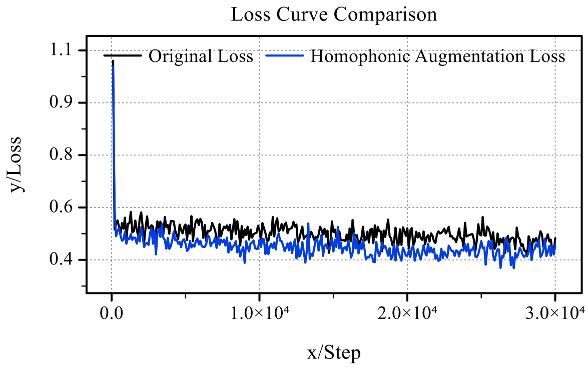

Figure 4: Long-term training loss curve.

Figure 3 and 4 show the short-term and long-term loss curves during training, respectively. Models trained with the homophone-enhanced dataset exhibit better convergence compared to the original across both time scales. As illustrated in Figure 3, the enhanced model's loss curve drops steadily with lower volatility, indicating rapid adaptation to homophonic interference. In Figure 4, the enhanced model consistently maintains a lower loss over long-term training, demonstrating improved learning ability under complex linguistic disturbances.

5.2.2 Model Test Performance Comparison

To verify the effect of homophone-aware fine-tuning, we compare model performance under equal training steps.

As seen in Figure 5, the model trained with homophone enhancement (red curve) significantly outperforms the original model (blue curve) early in training. While both improve over time, the enhanced model consistently maintains higher accuracy, validating its superior capacity in detecting homophone variations.

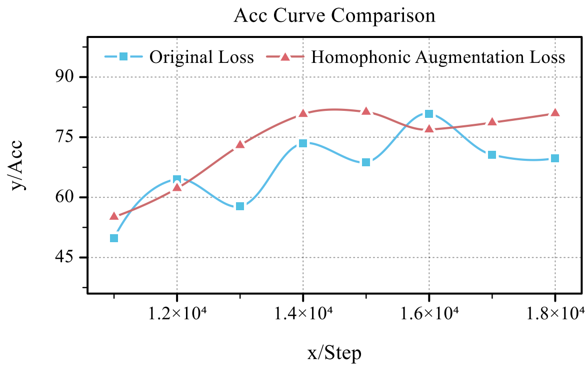

Figure 5: Accuracy comparison of original vs. homophone-enhanced models on the test set.

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 2, the baseline models exhibit substantial performance differences between 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 0.6531 to 0.8759 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 2.7 and 2.1 percentage points, respectively. 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,

Table 2: Model performance comparison

Models	COLD Test				HED - COLD Test			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Qwen2.5-3B	0.5656	0.4763	**0.9682**	0.6385	0.5259	0.4540	**0.9753**	0.6196
Qwen2.5-3B+ours	0.8232	0.8894	0.8078	0.8467	0.8529	0.9041	0.8364	0.8689
Qwen2.5-7B	0.7366	0.6501	0.7247	0.6854	0.7221	0.6453	0.6610	0.6531
Qwen2.5-7B+ours	0.8279	**0.8912**	0.8111	**0.8493**	**0.8587**	**0.9121**	0.8425	**0.8759**
Bert	0.8144	0.7246	0.8667	0.7893	0.8082	0.7247	0.8310	0.7742
Bert+ours	0.8212	0.7336	0.8605	0.7920	0.8290	0.8008	0.8018	0.8013
Chinese-roberta-wwm-ext	0.8251	0.7379	0.8657	0.7967	0.8136	0.7409	0.8134	0.7755
Chinese-roberta-wwm-ext+ours	**0.8371**	0.8012	0.7826	0.7918	0.8364	0.7852	0.8072	0.7961

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.

5.2.3 Homophone Adaptability Analysis

To assess the impact of homophone data, we calculate F1-score differences between COLD and HED-COLD test sets:

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

Table 3: F1-score difference across test sets

Models	Gender	Region	Race	Total
Qwen2.5-3B	-0.026	-0.036	0.007	-0.019
Qwen2.5-3B+ours	0.024	0.017	0.029	0.022
Qwen2.5-7B	-0.063	-0.049	0.006	-0.032
Qwen2.5-7B+ours	0.025	0.020	0.029	0.025
Bert	-0.021	-0.012	-0.005	-0.0151
Bert+ours	0.023	0.015	0.010	0.0093
Chinese-roberta-wwm-ext	-0.013	-0.032	-0.022	-0.0212
Chinese-roberta-wwm-ext+ours	0.007	0.008	0.002	0.0043

As shown in Table 3, baseline models without homophone augmentation exhibit notable performance degradation on the HED-COLD test set compared to the original COLD set, with F1-score reductions observed across multiple task categories. The most pronounced drops occur in the *Region* and *Gender* categories. For instance, Qwen2.5-7B shows an F1-score decline of 0.049 in Region and 0.063 in Gender, indicating a lack of robustness in handling phonetic perturbations within these contexts. In contrast, models fine-tuned with our homophone-augmented data demonstrate consistent performance gains across all categories, with the most stable and significant improvements observed in the *Race* category. These results suggest that the proposed augmentation strategy not only improves overall model robustness but also mitigates sensitivity disparities across task-specific categories.

A deeper investigation reveals that *Gender* and *Region* are the categories most susceptible to phonetic attacks, largely due to their lexical characteristics. Terms related to gender and geographical regions are frequently manipulated via homophonic substitutions to evade detection—for example, replacing ‘‘东北’’ (northeast) with ‘‘东百’’ or ‘‘男人’’ (man) with ‘‘蝻人.’’ Such transformations preserve phonetic similarity while altering surface forms, making them difficult for character-level models to detect. Our proposed homophone-enhancement strategy addresses this challenge by incorporating structured homophonic variants during training. The results underscore the necessity of modeling phonetic variation in Chinese safety-sensitive NLP tasks, especially when defending against adversarial attacks targeting social attributes.

5.2.4 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 large language models (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

Table 4: Models' performance on the ToxiCloakCN

Models	Training Set	Instruction Type	Homophone	Base
COLDetector	COLD	-	0.566	0.625
	HED-COLD	-	0.658	0.647
LLAMA-3-8B	COLD	Chinese_text	0.599	0.689
	HED-COLD	Chinese_text	0.702	0.693
Mistral	COLD	Chinese_text	0.547	0.691
	HED-COLD	Chinese_text	0.718	0.704
Qwen1.5-MoE A2.7B	COLD	Chinese_text	0.650	0.700
	HED-COLD	Chinese_text	0.719	0.712
Qwen2.5-3B	COLD	Chinese_text	0.603	0.688
	HED-COLD	Chinese_text	0.705	0.697
Qwen2.5-7B	COLD	Chinese_text	0.624	0.693
	HED-COLD	Chinese_text	0.725	0.701

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 4, 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 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

This study addresses the challenge of defending against homophonic adversarial attacks in Chinese online environments by proposing a robustness-enhancing framework for large language models. We introduce HED-COLD dataset and develop a homophone-aware pretraining strategy to equip models with phonetic resilience. Experimental results consistently show that traditional models suffer from significant performance degradation under homophonic attack scenarios, whereas models trained with our augmented data demonstrate improved stability and robustness. In particular, the proposed method achieves balanced improvements across sensitive attributes such as gender and region, highlighting its domain-generalizable effectiveness. Furthermore, evaluation on the out-of-domain ToxiCloakCN benchmark confirms that our strategy not only enhances detection of phonetic adversaries but also improves performance on clean inputs, validating its broad transferability and real-world applicability.

In future work, we plan to explore multimodal homophone attacks that combine phonetic perturbations with visual and structural noise, such as emoji insertion, character distortion, and code-switching. Finally, we envision building adaptive adversarial training pipelines that integrate phonological knowledge dynamically during pretraining and finetuning, enabling more robust and context-aware defense systems for open-domain Chinese NLP applications.

7 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.

8 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. 2023. Model-agnostic meta-learning for multilingual hate speech detection. *Preprint*, arXiv:2303.02513.

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. 2020. N-ltp: An open-source neural language technology platform for chinese. *arXiv preprint arXiv:2009.11616*. Accepted to appear in EMNLP 2021 (Demo).

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. *Preprint*, arXiv:1908.06083.

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.

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, Chao Min, Liang Yang, and Hongfei Lin. 2023. Facilitating fine-grained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. *arXiv preprint*, abs/2305.04446.

David Noever. 2018. Machine learning suites for online toxicity detection. *Preprint*, arXiv:1810.01869.

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.

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. *Preprint*, arXiv:2406.12223.

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. *Preprint*, arXiv:2204.10521.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023. Cross-cultural transfer learning for chinese offensive language detection. *Preprint*, arXiv:2303.17927.

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 epochs with a relatively low learning rate of 1×10^{-8} to ensure stable fine-tuning. 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.

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。	高，脯权为了反粄拳真是什么恶都能作。
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?	这么说冻倍还人人黑涩会呢，你也是从知乎上了解的吧?
race	1	印度人打架是不敢还手的，欺软怕硬	印肚人欺软怕硬，打价是不敢还手的
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"?	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"?
gender	0	我们这些偏瘦的男士，一直再为120130的体重奋斗。	我们这些偏瘦的蝻士，一直再为120130的体重粪逗。
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。	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。
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……	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……
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。	林书豪，作为一个哑裔很历志但是这次的伤非同小可，真的完了。
…	…	…	…

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。

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。
2.判断规则：
" - 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。"
" - 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。"
" - 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。"
3. 待检测言论："这个飞舞糊蛹人简直是一头春竹，怎么教都不会" ← user input
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。如果言论包含歧视或冒犯性的仇恨言论，请输出"1"；否则，请输出"0"。"

Assistant:
1

Figure 7: dialogue example of offensive language detection.