

Rule-Based Perturbation, Real Evaluation: How Far Do Fine-Tuned Encoders Go on Chinese Phonetically Cloaked Toxic Language?

Minggui Duan

DIT

Università di Bologna

Viale Filippo Corridoni 20, 47121 Forlì, FC, Italy

minggui.duan2@unibo.it

Shijie Liu

College of Language and Culture

Northwest A&F University

Shaanxi 712100, China

shijie.liu@nwafu.edu.cn

Abstract

Warning: this paper contains offensive language.

Phonetic cloaking, the deliberate use of homophones, Pinyin, and numeric substitutions to disguise toxic content, poses a persistent challenge to Chinese toxic language detection. Existing work mainly relies on rule-based synthetic perturbations for benchmark evaluation, leaving uncertain how well such results transfer to naturally occurring cloaking. We present a controlled evaluation of fine-tuned encoders on PCR-ToxiCN, a dataset of naturally occurring, phonetically cloaked Chinese posts. Fine-tuning ChineseBERT and RoBERTa-wwm-ext on the CNTP synthetic perturbation dataset, we quantify synthetic-to-real performance gaps across four cloaking strategies. Synthetic perturbation training produces consistent but limited improvements, with diminishing returns as perturbation coverage expands. Hanzi (Chinese-character) replacement remains consistently difficult for both encoders, and few-shot fine-tuning on roughly 50 real examples provides only marginal improvement. The few-shot experiment, however, cannot distinguish representation limits from data scarcity or sampling effects. Overall, our findings caution against equating synthetic benchmark performance with real-world robustness.

1 Introduction

Social media enables large-scale public expression, but it also exposes users to toxic language, harassment, and cyberbullying. In 2023, 23-year-old Chinese woman Linghua Zheng committed suicide after enduring six months of intense stigmatization and cyberbullying merely for posting a photo of her pink hair online (De Pacina, 2023). The case illustrates the potential harms of toxic online environments and the need for effective moderation. Consequently, Natural Language Processing (NLP) research has increasingly addressed auto-

matic toxic-language detection across languages, including Vietnamese (Nguyen et al., 2021), Chinese (Lu et al., 2023), Arabic (Mubarak et al., 2020), Korean (Jeong et al., 2022; Park et al., 2023), Turkish (Çöltekin, 2020), and Danish (Sigurbergsson and Derczynski, 2020).

Automatic detection of toxic language is central to content moderation at scale (Nobata et al., 2016). In the Chinese online environment, this task is persistently undermined by *phonetic cloaking*: users substitute offensive characters with homophones, Pinyin spellings, or phonetically similar digits, producing text that evades lexical filters while remaining perfectly understandable to human readers (Xiao et al., 2024; Guo et al., 2025).

The research community has focused primarily on *synthetic perturbation*. Datasets such as ToxiCloakCN (Xiao et al., 2024) and CNTP (Yang et al., 2025) apply rule-based transformations (homophone swaps, Pinyin conversion, emoji substitution, etc.) to existing toxic corpora, generating large-scale training and evaluation data. Following prior work (Xiao et al., 2024; Guo et al., 2025), we refer to such rule-generated perturbations as *synthetic*, in contrast to naturally occurring cloaking produced by real users. Though this method is scalable and reproducible, it carries a structural weakness. Models are both trained and evaluated on perturbations generated by the same family of rules. Whether performance on such benchmarks reflects robustness to the cloaking that real users produce has remained an open question.

The recent release of PCR-ToxiCN (Guo et al., 2025), a dataset of 500 naturally occurring, phonetically cloaked posts collected from the Chinese social platform *Xiaohongshu* (RedNote), makes this question empirically answerable. Its authors benchmarked state-of-the-art LLMs and found the best model (o3-mini) reaching 0.672 positive-class F_1 on the balanced binary offensive/non-offensive task. However, while recent benchmarks increas-

ingly focus on evaluating large language models (Xiao et al., 2024), the behavior of fine-tuned encoders, which are standard supervised baselines in this domain, remains under-explored under real-world cloaking.

Thus, we fine-tune two representative Chinese encoders, ChineseBERT (Sun et al., 2021) and RoBERTa-wwm-ext (Cui et al., 2021), on the CNTP (Yang et al., 2025) synthetic perturbation dataset, and evaluate them on PCR-ToxicCN across its four cloaking strategies: Hanzi (Chinese-character) replacement (HR), alphabet replacement (AR), numeric replacement (NR), and mixed replacement (MR). Our paper aims to address three questions:

- **RQ1:** How well do encoders trained on unperturbed toxic text transfer to real-world phonetic cloaking?
- **RQ2:** Does expanding the coverage of synthetic perturbation types systematically improve real-world transfer?
- **RQ3:** What are the empirical gains of introducing a minor portion of naturally occurring cloaked data during fine-tuning?

Our findings are largely negative. For RoBERTa, adding CNTP’s homophone-substitution augmentation captures nearly all of the gain from synthetic perturbation training, while adding the seven remaining CNTP types—Pinyin initials, full Pinyin, emoji substitution, character decomposition, radical substitution, character swapping, and traditional-script conversion—yields no consistent improvement. ChineseBERT improves more gradually but never surpasses RoBERTa. HR, the most frequent strategy in PCR-ToxicCN, yields the lowest detection F_1 under most configurations. Augmenting training with a stratified 10% sample of real cloaked posts (approximately 50 examples) yields only a marginal gain (at most +0.03 overall F_1). This pattern is consistent with a remaining representation mismatch, but it does not rule out data scarcity or dataset-sampling effects.

2 Background

Phonetic cloaking taxonomy. We adopt the four-way surface-form taxonomy of Guo et al. (2025). **HR** (Hanzi replacement) substitutes an offensive character with a homophonous or near-homophonous character. **AR** (alphabet replacement) rewrites characters in Latin script, ranging

CNTP type	PCR strategy	Coverage
Homophone	HR	partial
Pinyin initials	AR	partial
Full Pinyin	AR	partial
Emoji, Decomp., Radical, Swap, Traditional	—	none
—	NR	none
—	MR	none

Table 1: Correspondence between CNTP synthetic perturbation types and PCR-ToxicCN real-world cloaking strategies.

from Pinyin initials and full Pinyin to English words with similar pronunciation. **NR** (numeric replacement) uses Arabic digits whose Chinese readings approximate the target syllable. **MR** (mixed replacement) combines two or more strategies within a single expression, requiring the reader to switch decoding modes mid-phrase.

Datasets. ToxicCN (Lu et al., 2023) is a benchmark of 12,011 Chinese posts collected from *Zhihu* and *Tieba* between June 2021 and November 2022 and annotated for toxicity across topics on gender, race, region, and LGBTQ. **CNTP** (Yang et al., 2025) derives from ToxicCN. It selects a subset of toxic and non-toxic sentences (the *base* set) and applies eight rule-based perturbations to the toxic sentences: homophone substitution, Pinyin initials, full Pinyin, emoji substitution, character decomposition, radical substitution, character swapping, and traditional-script conversion. **PCR-ToxicCN** (Guo et al., 2025) is independently curated from RedNote rather than sampled from ToxicCN. Its authors manually collected candidate posts under the four-way PCR taxonomy, filtered noisy or context-dependent cases, and retained a balanced set of 500 posts (250 offensive, 250 non-offensive): HR (352), AR (87), NR (32), and MR (29).

Coverage correspondence. The synthetic and real taxonomies align only partially (Table 1). CNTP’s homophone perturbation corresponds to HR, and its two Pinyin perturbations to AR. Real AR, however, also includes English near-homophones and placeholder letters that no CNTP rule generates. NR and MR have no synthetic counterpart at all. This asymmetry is central to our analysis (Section 6).

3 Related Work

Chinese toxic language detection. COLD (Deng et al., 2022) established an early large-scale

benchmark. ToxiCN (Lu et al., 2023) extended coverage to implicit toxicity with a hierarchical taxonomy. Fine-tuned encoders, particularly RoBERTa-wwm-ext (Cui et al., 2021), are standard supervised baselines throughout this literature. ChineseBERT (Sun et al., 2021) incorporates Pinyin and glyph embeddings into pre-training, making it a natural candidate for testing whether explicit pronunciation features improve robustness to phonetic cloaking. Offensive-language detection has also been studied in other non-Latin-script languages. For Korean, benchmark studies show that pretrained encoders can perform effectively on general offensiveness detection, while more implicit and fine-grained offensive expressions remain substantially more challenging (Jeong et al., 2022; Park et al., 2023). Chinese phonetic cloaking introduces a different source of difficulty: the intended offensive expression is preserved while its surface form is deliberately altered.

Robustness to cloaking perturbations. ToxiCloakCN (Xiao et al., 2024) demonstrated that homophone and emoji perturbations sharply degrade detection performance. CNTP (Yang et al., 2025) expanded rule-based perturbation coverage to eight types and benchmarked LLMs under zero-shot, in-context, and supervised settings. Both works generate their evaluation data with rules, leaving open how models fare against human-crafted cloaking.

Real-world evaluation. Guo et al. (2025) introduced PCR-ToxiCN and showed that state-of-the-art LLMs, including reasoning models, perform poorly on it, with chain-of-thought prompting further degrading results. Their evaluation covers only prompted LLMs, while ours is complementary, covering the fine-tuned encoder paradigm that dominates prior supervised work.

4 Experimental Setup

4.1 Data

All experiments draw on a unified pool built from CNTP and PCR-ToxiCN. Training data comes exclusively from CNTP, while PCR-ToxiCN is reserved for evaluation, except in the few-shot condition of §5.3.

CNTP base. The base set contains 2,533 toxic and 2,696 non-toxic sentences drawn from ToxiCN. It provides the fundamental toxic/non-toxic signal in every training configuration. Perturbation sets

are always added on top of it, never used alone, because CNTP generates perturbed variants only from toxic source sentences and therefore contains no synthetic non-toxic perturbation examples. Importantly, phonetic replacement is not itself a toxicity label: PCR-ToxiCN contains non-offensive PCR instances as well (Guo et al., 2025). We discuss the implications of this mismatch further in the [Limitations section](#).

CNTP perturbations. Each of the eight perturbation types contains roughly 2,500 perturbed variants of the toxic base sentences (all labeled toxic). We align every perturbed variant back to its source sentence via ToxiCN sentence identifiers, recovering alignment through exact source-text matching where identifiers were inconsistent across files.

Splits. We split at the level of source-sentence identifiers, stratified by label (70/10/20 train/validation/test, seed 42). All perturbed variants inherit the split of their source sentence, so no sentence appears in different splits under different surface forms, eliminating leakage between training and validation.

PCR-ToxiCN. For RQ1 and RQ2, all 500 PCR-ToxiCN instances are held out for evaluation and reported both overall and per strategy. RQ3 is the only exception: a stratified 10% is added to training and evaluation uses the remaining 90%.

4.2 Models and Training

We fine-tune **ChineseBERT-base** (Sun et al., 2021), whose Pinyin embeddings make it the natural candidate for phonetic robustness, and **RoBERTa-wwm-ext** (Cui et al., 2021), the standard baseline in prior work. Both receive a randomly initialized binary classification head.

Training uses AdamW (learning rate 2×10^{-5} , weight decay 0.01), batch size 32, maximum sequence length 128, and up to 10 epochs with early stopping (patience 3) on validation F_1 . Because perturbation sets contain only toxic samples, stacking multiple types skews the label distribution, so we use a weighted random sampler to approximately balance class sampling during training. All results are averaged over five random seeds and reported as mean $\pm$ standard deviation.

4.3 Evaluation

Following Guo et al. (2025), we report positive-class (toxic) F_1 as the primary metric, along-

Hyperparameter	Value
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.01
Batch size	32
Maximum length	128
Maximum epochs	10
Early stopping	patience 3
Random seeds	5

Table 2: Training settings used for both encoders.

side precision and recall. This makes our numbers directly comparable to their balanced binary offensive/non-offensive LLM benchmark, where the best model reaches 0.672 F_1 . For key comparisons, we additionally report two-sided paired t -tests over the five seed-level scores.

5 Results

5.1 RQ1: Zero-Perturbation Baseline

We first train both encoders on the CNTP base alone (clean toxic and non-toxic text, no perturbations) and evaluate on PCR-ToxicCN (Table 3, top block).

Both models transfer poorly. RoBERTa reaches 0.531 overall F_1 and ChineseBERT 0.509, far below their in-domain validation performance (approximately 0.90). The per-strategy breakdown is revealing: RoBERTa handles AR comparatively well (0.795) but collapses on HR (0.415), the most frequent real-world strategy. Across configurations, precision consistently exceeds recall by a wide margin (e.g., 0.665 vs. 0.412 overall for RoBERTa at seed 42), meaning cloaked toxic posts are overwhelmingly misclassified as non-toxic, the same false-negative failure mode Guo et al. (2025) observed for LLMs.

Contrary to expectation, ChineseBERT’s Pinyin embeddings confer no advantage: it trails RoBERTa on every strategy, most severely on AR (0.584 vs. 0.795). We return to this in Section 6.

5.2 RQ2: Perturbation Coverage Ablation

We next add perturbation types incrementally, ordered by their correspondence to real strategies: homophone (matching HR), then Pinyin initials and full Pinyin (matching AR), then the five remaining types with no real-world counterpart.

The results (Figure 1; Table 3, middle blocks) show a sharp saturation pattern. For RoBERTa, the single homophone perturbation delivers nearly the

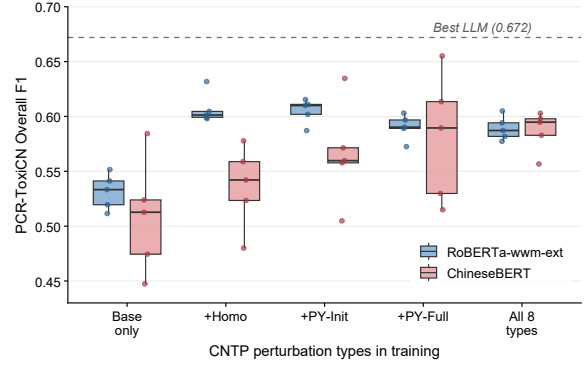


Figure 1: Overall positive-class F_1 on PCR-ToxicCN as CNTP perturbation types are added incrementally. Boxes show the distribution over five seeds, and the dashed line marks the best prompted-LLM result from Guo et al. (2025).

entire benefit: overall F_1 rises from 0.531 to 0.607, with HR improving from 0.415 to 0.533. NR and MR also improve despite having no matched training signal, which may indicate that homophone augmentation teaches a weak general cue that “surface form may deviate from intent.” Every subsequent addition yields flat or slightly negative changes. The full eight-type setting (0.589) is statistically indistinguishable from, and nominally below, the one-type configuration (paired t -test, $p = 0.08$), whereas the initial jump from base to one type is highly significant ($p < 0.001$). ChineseBERT improves more gradually (0.509 $\rightarrow$ 0.587) but never overtakes RoBERTa.

Two observations stand out. First, adding perturbation types with no direct real-world counterpart provides no consistent additional benefit, especially for RoBERTa, supporting the practical relevance of the correspondence asymmetry in Table 1. Second, even the best-matched perturbation leaves HR far below AR: rule-generated homophones appear to capture only a narrow slice of the near-homophone spectrum represented in real data.

5.3 RQ3: Few-Shot Real Data

Finally, we ask how much a small quantity of genuine cloaked data helps. We move a stratified 10% of PCR-ToxicCN (approximately 50 instances) into training alongside all eight CNTP perturbation types, evaluating on the remaining 90%.

The gain is marginal: +0.026 overall F_1 for RoBERTa (0.589 $\rightarrow$ 0.615; paired t -test, $p = 0.02$) and a non-significant +0.018 for ChineseBERT ($p = 0.36$). Even with authentic examples of every cloaking strategy present in training, neither

Training config.	Model	Overall	HR	AR	NR	MR
Base only (RQ1)	ChineseBERT	.509±.052	.476±.073	.584±.018	.471±.066	.604±.061
	RoBERTa	.531±.016	.415±.026	.795±.017	.643±.024	.670±.027
+ 1 type (Homo)	ChineseBERT	.536±.037	.508±.054	.628±.046	.487±.042	.541±.044
	RoBERTa	.607±.014	.533±.022	.782±.026	.718±.042	.718±.024
+ 2 types (+PY-Init)	ChineseBERT	.566±.046	.489±.090	.708±.066	.499±.077	.734±.079
	RoBERTa	.605±.011	.522±.015	.776±.011	.712±.054	.707±.028
+ 3 types (+PY-Full)	ChineseBERT	.581±.058	.498±.094	.739±.024	.485±.102	.747±.025
	RoBERTa	.590±.011	.498±.009	.778±.014	.695±.029	.700±.012
+ 8 types (all)	ChineseBERT	.587±.019	.500±.020	.747±.019	.568±.045	.752±.014
	RoBERTa	.589±.011	.504±.013	.758±.019	.687±.032	.715±.011
+ Few-shot real (RQ3)	ChineseBERT	.605±.033	.532±.060	.738±.019	.573±.056	.755±.003
	RoBERTa	.615±.011	.539±.014	.779±.015	.707±.052	.712±.020
Best prompted LLM (Guo et al., 2025)		.672	—	—	—	—

Table 3: Positive-class F_1 (mean $\pm$ std over 5 seeds) on PCR-ToxicCN, overall and per cloaking strategy. Configurations add CNTP perturbation types incrementally on top of the CNTP base. The few-shot block adds a stratified 10% of PCR-ToxicCN to training (evaluation on the remaining 90%). Best encoder result per column in **bold**.

encoder approaches the best LLM result of 0.672, and HR remains near 0.53.

Approximately 50 training examples cannot cover the combinatorial space of real cloaking, and the stratified subset is too small to determine how performance would scale with hundreds or thousands of authentic examples. The limited gain therefore shows only that small-scale in-domain augmentation is insufficient in our setting. It is consistent with a representation mismatch, but does not rule out data scarcity, collection bias, or train–test distribution effects as alternative explanations.

6 Analysis

6.1 Why HR Is Systematically Hard

HR is the most frequent strategy in PCR-ToxicCN (70%) and is consistently among the hardest conditions (Figure 2). One plausible contributor is the *near-homophone spectrum* (Guo et al., 2025): near-homophones are substantially more common than perfect homophones in PCR-ToxicCN, including substitutions that differ in tone or other phonetic components. Rule-based generation, in contrast, relies on predefined transformation rules. Training on such data may encourage memorization of lexical correspondences without guaranteeing a general notion of phonological similarity. When a real post uses only an approximate phonetic neighbor, those learned correspondences may therefore fail to transfer.

6.2 The Pinyin Embedding Paradox

ChineseBERT’s Pinyin channel should, in principle, make phonetic information more accessible. In practice, it underperforms RoBERTa across the board (Figure 2, right panel). We conjecture two contributing factors. First, ChineseBERT converges noticeably more slowly in our setting (validation F_1 plateaus near 0.79 vs. 0.90 for RoBERTa), which may reflect greater data requirements for its fused character-Pinyin-glyph input. Second, the Pinyin embedding encodes the pronunciation of the *written* character; in a cloaked post, that is the substitute rather than the intended toxic form. The Pinyin channel helps only if the model also learns relations among similar-sounding syllables. Our results therefore suggest that the available phonetic representation may be underutilized in this fine-tuning setting, rather than proving that it is intrinsically ineffective.

6.3 Error Patterns

Following WOA data policy, we illustrate error patterns with synthetic reconstructions that preserve the structure of real failures without reproducing user content. Offensive morphemes are masked throughout.

Near-homophone HR (missed). A post replaces one character of a two-character insult with a same-initial, different-tone everyday character, roughly the equivalent of writing “s*illy” with an innocuous word that merely sounds similar. Both models read the surface form literally and predict non-toxic.

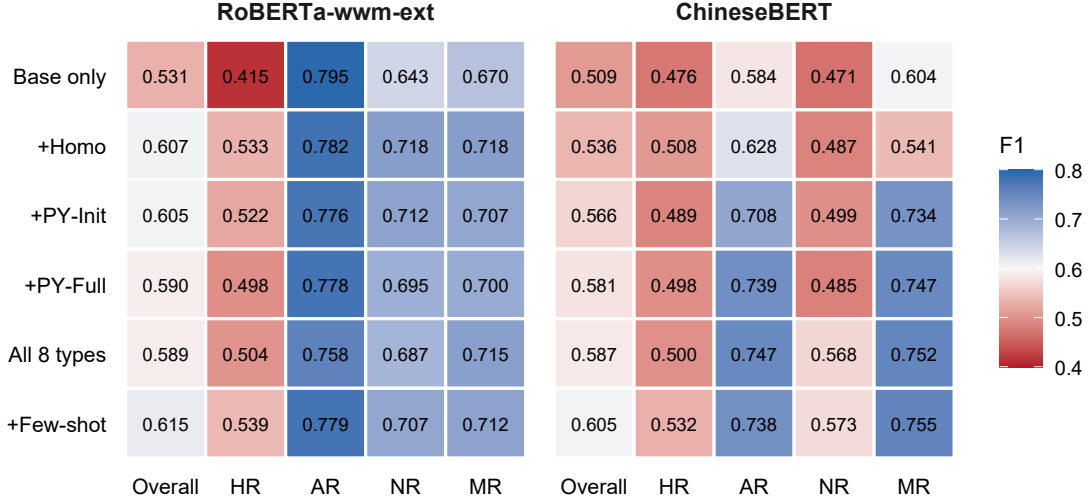


Figure 2: Positive-class F1 (mean over five seeds) across training configurations and PCR-ToxicCN cloaking strategies. The persistently red HR column illustrates its systematic difficulty, and the near-constant rows below +Homo for RoBERTa visualize the saturation effect.

The substitution is invisible without phonological reasoning.

English-homophone AR (missed). A post embeds an English word whose pronunciation approximates a Chinese vulgarity. No CNTP rule produces English near-homophones, and both models miss such cases, illustrating that real AR is broader than its synthetic counterpart.

Compositional MR. A post combines a digit substitution with a homophone inside one two-character expression, requiring two decoding strategies simultaneously. Models frequently fail on the cloaked span, yet MR nevertheless scores above HR in Table 3, which we attribute to MR posts carrying more residual toxic context outside the cloaked span.

6.4 Encoders vs. LLMs

Our best encoder configuration (0.615) trails the best prompted LLM (0.672), and no configuration closes the gap. Combined with the finding of Guo et al. (2025) that LLMs themselves perform poorly and that Pinyin-augmented prompting recovers part of the loss, the results are consistent with a need for stronger handling of phonological relations. Our experiments do not, however, isolate representation limits from data scarcity or sampling effects. For encoders specifically, RQ3 shows that approximately 50 real examples are insufficient to close the gap in our setting.

7 Conclusion

We presented a controlled evaluation of fine-tuned encoders on real-world Chinese phonetic cloaking, using CNTP for synthetic training and PCR-ToxicCN for evaluation. Synthetic perturbation training helps, but its gains saturate quickly for RoBERTa and remain limited overall. Hanzi replacement remains consistently difficult, Pinyin-aware pre-training provides no measurable advantage in our setting, and adding approximately 50 real examples yields only marginal improvement. Together with Guo et al. (2025), our results indicate that rule-based synthetic benchmarks can overstate robustness to naturally occurring cloaking. Evaluating on real user-produced data is therefore important for measuring progress, while larger and more representative real-data studies are needed to determine how much of the remaining gap is attributable to model representation versus data and sampling.

Limitations

Our study inherits the scale limitations of PCR-ToxicCN: with only 32 NR and 29 MR instances, per-strategy results for these categories carry wide uncertainty, and we caution against strong conclusions from their point estimates. The few-shot condition of RQ3 uses approximately 50 real training instances; this is sufficient only to test a small-data regime and does not justify concluding that data quantity is irrelevant. Larger real-data budgets

could yield substantially different results.

PCR-ToxicCN is independently curated from Red-Note and deliberately balanced to 250 offensive and 250 non-offensive instances (Guo et al., 2025). Its taxonomy-guided collection and balancing make it suitable as a targeted benchmark, but they may also introduce sampling or distributional biases. Our experiments do not isolate such effects from the difficulty of real-world perturbations. In addition, phonetic replacement can occur in non-offensive language, whereas CNTP generates synthetic perturbations only from toxic source sentences. Consequently, models may partially exploit correlations between “being perturbed” and toxicity that would not hold in unrestricted social-media data.

Our conclusions concern rule-based synthetic perturbation; LLM-generated adversarial cloaking may transfer differently and merits separate investigation. Finally, we evaluate only ChineseBERT-base and RoBERTa-wwm-ext. Future work should test additional Chinese encoders and larger variants such as ChineseBERT-large before generalizing the architectural conclusions beyond the two models studied here.

Ethical Considerations

This work uses only pre-existing, publicly released research datasets (ToxicCN, CNTP, PCR-ToxicCN) under their respective licenses. We collected no new user data. Illustrative examples in §6 are synthetic reconstructions that preserve linguistic structure without reproducing any user’s post, and offensive morphemes are masked, in line with WOAHI data policies. Detection research on cloaked toxicity is dual-use: analyses of model blind spots could in principle inform evasion. We judge the marginal risk low, since the cloaking strategies discussed are already publicly documented in the source datasets, while the moderation benefit of understanding these blind spots is substantial.

Acknowledgments

The research of M. Duan is supported by the China Scholarship Council (CSC) under Grant No. 202508310058. The research of S. Liu is supported by the Fundamental Research Funds for the Central Universities, Chinese Universities Scientific Fund, under Grant No. XYTD2026-21.

References

- Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6180.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.
- Michelle De Pacina. 2023. [Suicide of pink-haired chinese woman sparks campaign to combat cyberbullying](#). Yahoo! News. Accessed: 2026-07-03.
- Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Haotian Guo, Jianfei He, Jiayuan Ma, Hongbin Na, Zimu Wang, Haiyang Zhang, Qi Chen, Wei Wang, Zijing Shi, Tao Shen, and Ling Chen. 2025. Lost in pronunciation: Detecting Chinese offensive language disguised by phonetic cloaking replacement. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2538–2550. Association for Computational Linguistics.
- Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. [KOLD: Korean offensive language dataset](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- 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.
- Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. [Overview of OSACT4 Arabic offensive language detection shared task](#). In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France. European Language Resources Association.
- Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Constructive and toxic speech detection for open-domain social media comments in vietnamese. In *International conference on industrial, engineering and other applications of applied intelligent systems*, pages 572–583. Springer.

- Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153.
- San-Hee Park, Kang-Min Kim, O-Joun Lee, Youjin Kang, Jaewon Lee, Su-Min Lee, and SangKeun Lee. 2023. “why do I feel offended?” - Korean dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1142–1153, Dubrovnik, Croatia. Association for Computational Linguistics.
- Gísli Sigurbergsson and Leon Derczynski. 2020. Offensive language detection: Danish dataset and Baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4717–4724.
- Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. Chinese-BERT: Chinese pretraining enhanced by glyph and Pinyin information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2065–2075. Association for Computational Linguistics.
- 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. EMNLP 2024.
- Shujian Yang, Shiyao Cui, Chuanrui Hu, Haicheng Wang, Tianwei Zhang, Minlie Huang, Jialiang Lu, and Han Qiu. 2025. Exploring multimodal challenges in toxic Chinese detection: Taxonomy, benchmark, and findings. *Preprint*, arXiv:2505.24341. Findings of ACL 2025.