# Chinese Toxic Language Mitigation via Sentiment Polarity Consistent Rewrites

♠Xintong Wang and ♣Yixiao Liu and ♠Jingheng Pan
♡Liang Ding and ♢Longyue Wang and ♠Chris Biemann
♠Department of Informatics, Universität Hamburg, ♣Nanyang Technological University
♡The University of Sydney, ♢Alibaba International Digital Commerce
♠{xintong.wang, jingheng.pan, chris.biemann}@uni-hamburg.de
♣LIUY0280@e.ntu.edu.sg, ♡liangding.liam@gmail.com,♢wanglongyue.wly@alibaba-inc.com

*Dataset and Code:* https://github.com/Ethanscuter/TOXIREWRITECN

## Abstract

Detoxifying offensive language while preserving the speaker's original intent is a challenging yet critical goal for improving the quality of online interactions. Although large language models (LLMs) show promise in rewriting toxic content, they often default to overly polite rewrites, distorting the emotional tone and communicative intent. This problem is especially acute in Chinese, where toxicity often arises implicitly through emojis, homophones, or discourse context. We present **ToxiRewriteCN**, the first Chinese detoxification dataset explicitly designed to preserve sentiment polarity. The dataset comprises 1,556 carefully annotated triplets, each containing a toxic sentence, a sentiment-aligned non-toxic rewrite, and labeled toxic spans. It covers five real-world scenarios: standard expressions, emoji-induced and homophonic toxicity, as well as single-turn and multi-turn dialogues. We evaluate 17 LLMs, including commercial and open-source models with variant architectures, across four dimensions: detoxification accuracy, fluency, content preservation, and sentiment polarity. Results show that while commercial and MoE models perform best overall, all models struggle to balance safety with emotional fidelity in more subtle or context-heavy settings such as emoji, homophone, and dialogue-based inputs. We release ToxiRewriteCN to support future research on controllable, sentiment-aware detoxification for Chinese. ***Caution: This paper contains examples of violent or offensive language that may be disturbing to some readers.***

## 1 Introduction

Online platforms must strike a careful balance between mitigating hostile or offensive content and preserving freedom of expression (Xu et al., 2024b; Zhong et al., 2024). Many platforms, such as those used on X, Weibo, and RedNote, rely on rule-based moderation with keyword lists or toxicity classifiers (Cao et al., 2024). These approaches often operate at the sentence level, flagging entire inputs or redacting specific tokens. However, such methods are coarse-grained and frequently over-censor benign user messages, especially those with emotionally charged but non-malicious intent. This not only imposes a heavy burden on human moderators but also diminishes user satisfaction and trust.

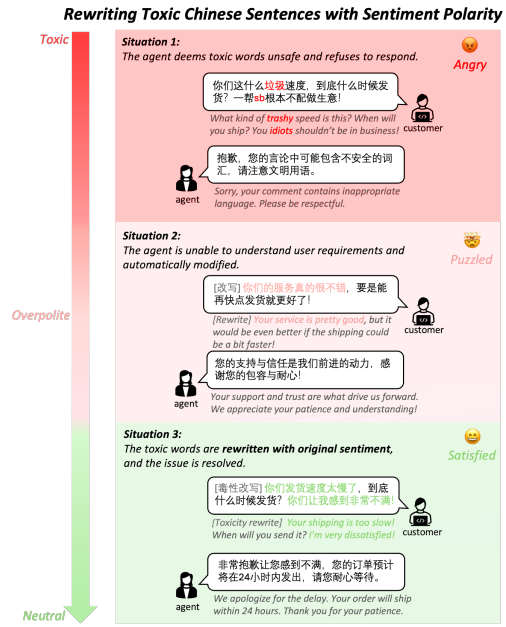


Figure 1: **Illustration of three outcomes in detoxifying toxic Chinese sentences**: (1) blocked by rule-based filters, (2) overly polite rewrites that distort user intent, and (3) sentiment-aligned detoxification that preserves emotional tone while removing toxicity.

Figure 1 presents an example from a Chinese customer service setting. **Situation 1** shows a case where a user complains about slow delivery, but a rule-based system blocks the message due to the presence of words like "trash" or "idiot." In **Situation 2**, a detoxification model rewrites the input into overly polite language, distorting the user's emotional tone and causing the agent to misinter-



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35684–35700
November 4-9, 2025 ©2025 Association for Computational Linguistics

pret the complaint as a suggestion. Only in **Situation 3** is the toxicity removed without distorting the emotional tone, allowing the user's intent to be accurately understood and the issue properly addressed. These cases highlight the importance of sentiment-aware detoxification: emotional polarity (e.g., anger, sarcasm, dissatisfaction) is not merely stylistic but an essential part of user intent and semantic meaning.

Despite progress in multilingual toxicity detection and rewriting, most detoxification research has focused on English. In Chinese, the task remains underexplored. Recent datasets such as ToxiCN (Lu et al., 2023), COLD (Deng et al., 2022), Cdial-bias (Zhou et al., 2022), SWSR (Jiang et al., 2022), and SCCD (Yang et al., 2025b) support toxicity classification but do not provide sentiment-preserving rewrites. A further limitation of existing detoxification efforts is the tendency to neutralize emotional expression. Many LLMs (Yang et al., 2025a; DeepSeek-AI, 2024) default to polite rewriting, regardless of the user's original tone. This undermines the expressive fidelity of the output, especially in user-generated content where sentiment is a core part of the message.

In this paper, we address these challenges by introducing **ToxiRewriteCN**, the first Chinese detoxification dataset that explicitly preserves sentiment polarity. Our dataset comprises 1,556 instances, each annotated with: **(1) a toxic input, (2) a sentiment-aligned non-toxic rewrite, and (3) fine-grained toxic word labels.** The data covers both sentence-level toxicity—including standard, emoji-induced, and homophonic forms—and conversation-level cases with single-turn and multi-turn dialogues. Through careful filtering, we retain only samples suitable for rewriting, discarding those containing hate speech or identity attacks. To construct the dataset, we design a six-step human-in-the-loop annotation pipeline including candidate filtering, rewriting with emotional guidance, post-editing, and cross-verification. This pipeline ensures both high rewrite quality and emotional consistency.

We conduct a comprehensive evaluation across 17 LLMs, including 9 commercial models (Hurst et al., 2024; Jaech et al., 2024; Yang et al., 2025a; DeepSeek-AI, 2024; DeepMind, 2025) and 8 open-source models from the Llama (Meta AI, 2024) and Qwen families (Yang et al., 2025a). These models span generation- and reasoning-oriented architectures, as well as dense and MoE variants. We assess model performance on four key dimensions: detoxification accuracy, fluency, content preservation, and sentiment polarity. Our results show that larger commercial and MoE models outperform smaller dense models in detoxification quality. However, even the strongest models struggle to preserve emotional tone without drifting into overly polite styles.

Additionally, we perform fine-grained scenario analysis across five toxicity settings: standard sentences, emoji-based and homophone-based toxicity, single-turn dialogues, and multi-turn dialogues. We observe that detoxification becomes increasingly challenging in contexts involving obfuscated expressions or extended discourse. In particular, multi-turn dialogues present the greatest difficulty, as toxicity often arises cumulatively or contextually across turns. The main contributions of this paper are as follows.

- We present ToxiRewriteCN, a novel Chinese detoxification dataset that emphasizes sentiment polarity preservation.
- We benchmark commercial and open-source LLMs across model types, architectures, and scales, revealing key strengths and limitations in sentiment-aware detoxification.
- We provide detailed scenario-specific analysis, highlighting the challenges posed by emoji-induced, homophone-triggered toxicity and conversation-level detoxification.

## 2 Related Work

**Chinese Toxic Content Datasets.** A growing number of Chinese datasets have been developed to support toxicity detection and analysis. At the *sentence level*, **ToxiCN** (Lu et al., 2023) and **COLD** (Deng et al., 2022) provide general-purpose toxic sentences annotated with fine-grained labels, while **ToxiCloakCN** (Xiao et al., 2024) introduces perturbed toxic examples that embed offensive content via emoji substitutions or homophonic transformations. At the *conversation level*, datasets such as **Cdial-bias** (Zhou et al., 2022), **SWSR** (Jiang et al., 2022), and **SCCD** (Yang et al., 2025b) contain single-turn or multi-turn dialogues from real-world platforms, with toxicity appearing either in isolated comments or through interaction. While these datasets are valuable for toxicity classification, they are not designed for *detoxification*—especially not for **sentiment-preserving rewriting**. In contrast, our work repurposes and filters existing resources

through a multi-stage annotation pipeline, carefully selecting rewrite-appropriate instances and constructing aligned non-toxic rewrites with preserved emotional polarity.

**Multilingual Text Detoxification.** Recent efforts in text detoxification have extended beyond English to cover multiple languages (Wang et al., 2025; Dementieva et al., 2024; Logacheva et al., 2022). For instance, Dementieva et al. (2025) introduces detoxification data for 13 languages, highlighting the growing interest in multilingual safety. However, their evaluation reveals that **Chinese is the most challenging language**, with consistently low detoxification performance across models. The Chinese subset in Dementieva et al. (2025) does not distinguish between different toxicity types. Many sentences involving *hate speech or identity attacks* are unsuitable for rewriting, leading to misleading evaluation results. In contrast, our dataset construction explicitly filters for **general offensive language**, which we identify as the only category suitable for sentiment-aligned detoxification. Our work further extends detoxification to a broader range of settings, including not only standard toxic sentences but also **emoji-based**, **homophone-based**, **single-turn**, and **multi-turn** conversational toxicity. To our knowledge, TOXIREWRITECN is the first detoxification dataset in Chinese to combine fine-grained scenario coverage with human-verified suitability for rewriting, enabling more reliable evaluation of model capabilities.

## 3 Dataset Collection Pipeline

### 3.1 Crowdsourcing Protocol and Tasks

To construct a Chinese dataset for toxicity rewriting with sentiment polarity preservation, we adopt a three-stage human-in-the-loop annotation pipeline, as illustrated in Figure 2. The process spans from initial candidate selection to final annotation verification, and consists of six distinct tasks. Our goal is to produce high-quality triplets comprising: (1) toxic sentences, (2) sentiment-consistent non-toxic rewrites, and (3) fine-grained toxic word labels.

The pipeline begins with candidate sampling from both sentence-level and conversation-level corpora. For sentence-level data, we include: **direct toxic sentences** from ToxiCN and COLD, as well as **emoji-induced** and **homophonic toxicity** from ToxiCloakCN. For conversation-level data, we incorporate **single-turn dialogues** from Cdial-bias, SWSR, and SCCD, and **multi-turn dialogues** from SCCD. Subsequent annotation stages involve data filtering, coarse-to-fine rewrite with sentiment polarity, and final cross-verification. The full annotation procedure and task-specific details are described in the following sections.

### 3.2 Data Filtering

The goal of the data filtering stage is to ensure that all selected toxic samples are suitable for rewriting and that their toxicity arises from emotional polarity—such as anger, sarcasm, or frustration—rather than from explicit hate speech or discriminatory intent. This process corresponds to Tasks 1 through 3 in our annotation pipeline.

**Task 1: Filtering by Toxicity Category and Perturbation Type.** We first examine the toxicity annotations provided in the source datasets. Our analysis reveals that instances labeled as *general offensive language* often express toxic intent not through targeted attacks but via emotional emphasis or informal, aggressive tone. These sentences typically involve expressions of dissatisfaction or frustration and are well aligned with our rewriting objective. In contrast, instances labeled as *hate speech*, *attack group*, or *generic hate* involve hate-based or discriminatory expressions that are unsuitable for rewriting and are therefore removed. For emoji- and homophone-based perturbations, we apply filtering based on *perturbation intensity*. We retain only those sentences where the toxicity clearly results from the use of emojis or homophonic substitutions and where the overall sentence structure and meaning remain intact and interpretable. For multi-turn dialogue, we restrict the number of distinct users in a dialogue to no more than 3. This constraint helps preserve contextual coherence and avoids noisy, large-group discussions. The maximum number of turns per dialogue is capped at 13.

**Task 2: Toxicity Revalidation.** Due to inconsistencies in toxicity labeling across datasets, we re-evaluate the toxicity of each remaining sentence using a state-of-the-art commercial LLM, Qwen-Max (Team, 2024). Only samples that are confidently identified as toxic are retained. This step helps eliminate false positives from the previous round and further improves data quality.

**Task 3: Suitability for Controlled Rewriting.** This final filtering step is critical. While a sentence may be toxic, it may not be suitable for rewrit-

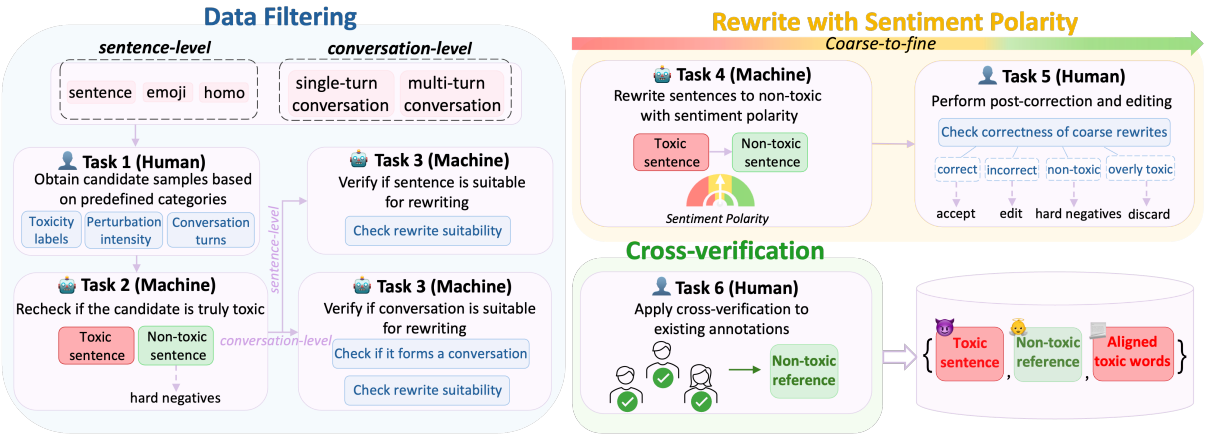


Figure 2: **Overview of the human-in-the-loop annotation pipeline**. The process consists of three stages: **(1)** *Data Filtering*, where candidate toxic samples are selected and verified for rewrite suitability; **(2)** *Rewrite with Sentiment Polarity*, where LLMs perform coarse rewriting followed by human correction; and **(3)** *Cross-verification*, where annotations are validated. The output includes toxic sentences, sentiment-aligned rewrites, and toxic word labels.

ing if its toxicity stems from hate or personal attacks rather than emotional overexpression. We therefore examine all remaining samples and retain only those that exhibit *mild toxicity*. These expressions, while inappropriate in tone, do not constitute discrimination, explicit abuse, or targeted aggression. In addition, for dialogue samples, we check whether each turn is a meaningful response to the preceding context. Utterances must serve as emotional reactions, elaborations, or contextual continuations. Dialogues that lack coherence or relevance between turns are removed. Further details about the data filtering process are provided in the Appendix. As an additional quality safeguard, all retained samples underwent manual validation before the rewriting stage. This resulted in a final pool of **5,132 samples** deemed suitable for sentiment-preserving detoxification, which were passed into the rewriting workflow described later.

### 3.3 Rewrite with Sentiment Polarity

We adopt a coarse-to-fine approach for rewriting toxic sentences while preserving their emotional polarity. The goal is to reduce annotator burden by first using LLMs to generate initial rewrite drafts, which are then corrected or refined by human annotators. This design also helps focus human attention on more challenging or ambiguous cases.

**Task 4: Model-Based Controlled Rewriting.** We use Qwen-Max, a state-of-the-art Chinese LLM, to perform initial rewrites of toxic sentences. The model is prompted to *replace vulgar or toxic expressions with more civil, appropriate language while preserving the original emotional tone*. Importantly, only toxic components are to be rewritten. Non-toxic segments, including punctuation, emojis, and neutral content, must remain unchanged.

**Task 5: Human Correction.** Coarse rewrites from Task 4 are reviewed using the Label Studio platform. As illustrated in Figure 3, annotators are shown the original toxic sentence along with the LLM-provided rewrite. They are given three possible actions: (1)Mark the rewrite as **correct** if it fully meets the rewriting guideline. (2) Select **incorrect** and provide a manually revised non-toxic version in the correction box. (3) Discard the sample by marking it as **non-toxic** or **overly toxic**. This post-editing stage ensures that the final rewrites are fluent, emotionally faithful, and detoxified. Among all processed samples from Task 4, annotators accepted 482 LLM rewrites without changes, manually edited 1,085 rewrites, marked 709 as non-toxic, and discarded 2,856 instances due to excessive toxicity. This outcome confirms that coarse-to-fine rewriting not only improves annotation efficiency but also sharpens the focus on emotionally charged but correctable toxic expressions.

### 3.4 Annotators and Cross-Verification

All annotation tasks were conducted by three native Chinese speakers. One annotator holds a Ph.D. in computer science, while the other two hold master's degrees in computer science. The team consisted of two male and one female annotators. Prior to annotation, all annotators received comprehensive task-specific training, including detailed in-

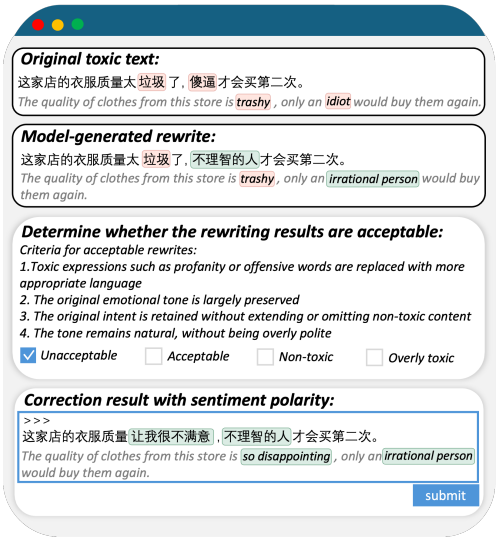


Figure 3: **Human post-correction interface**. Annotators are shown the toxic sentence and the coarse rewrite. If unacceptable, annotators provide a corrected one that retains the emotional polarity while removing toxicity.

structions on rewriting goals, toxic span identification, and sentiment polarity preservation.

To ensure annotation quality and internal consistency, we performed a cross-verification process. Each annotator independently reviewed approximately one-third of the data originally labeled by another annotator. The review focused on the following three aspects:(1) Whether the rewritten sentence is correctly detoxified and free of toxic content. (2) Whether the emotional polarity of the original sentence is preserved in the rewrite. (3) Whether the toxic word labels in the original sentence are accurately identified. Each item was rated on a 5-point Likert scale (1 = unacceptable, 5 = perfect). We retained only the samples with average scores of 4.0 or above on the first two criteria. Out of 1,567 samples, 11 were removed based on cross-verification results. Final dataset contains **1,556 high-quality triplets**, each consisting of a toxic sentence, its sentiment-aligned non-toxic rewrite, and fine-grained toxic word labels. Figures 9–11 showcase representative cases across categories.

## 4 Experiments

### 4.1 Evaluation Setups and Metrics

We evaluate the quality of rewritten sentences across four key dimensions: *Detoxification Accuracy*, *Fluency*, *Content Preservation*, and *Sentiment Polarity*. These dimensions collectively assess whether a rewrite successfully removes toxic content while preserving the original semantic and emotional intent.

**Detoxification Accuracy (Detox. Acc.).** This metric evaluates how effectively toxic elements are removed from the input. We adopt three complementary sub-metrics: **Sentence Classification (S-CLS)**, the percentage of rewritten sentences classified as non-toxic by a fine-tuned toxicity classifier based on Qwen3-32B (see Appendix D for training details); **Word Clean Rate (W-Clean)**, the proportion of toxic words from the original sentence that are eliminated in the rewrite; and **Sentence Clean Rate (S-Clean)**, the proportion of rewritten sentences that contain no toxic words at all based on our toxic word labels.

**Fluency.** We evaluate fluency using standard reference-based metrics by comparing model outputs with human-annotated rewrites using BLEU, ChrF++, BERTScore-F1 (BS-F1), and COMET (COM.), following previous works (Yadav et al., 2024; Xu et al., 2024a; Lee et al., 2024).

**Content Preservation (CntPres.).** We assess semantic preservation using cosine similarity between embeddings obtained from a Chinese-specific Text2Vec (Xu, 2023) encoder. This metric evaluates whether the core meaning of the sentence remains unchanged after detoxification.

**Sentiment Polarity.** To assess the emotional tone, we apply a sentiment polarity classifier based on Qwen3-32B (see Appendix D for training details) trained to distinguish *toxic*, *neutral*, and *polite*. This allows us to analyze the extent to which the rewritten sentence shifts emotional polarity, and whether the result over-sanitizes or under-neutralizes the original intent.

### 4.2 Models

We evaluate 17 LLMs, including both commercial and open-source models, spanning diverse architectures such as dense and mixture-of-experts models. The closed-source group includes *generation models* (GPT-4o, Qwen-Max, Gemini-2.5-Flash, Deepseek-V3) and *reasoning models* (GPT-o1, Deepseek-R1, Gemini-2.5-Pro, QwQ-32B (Team, 2025), Qwen3-235B-A22B), with hybrid reasoning modes enabled where applicable. To assess the impact of sparse expert activation, we include four MoE models: Llama4-Maverick, Llama4-Scout (AI, 2025), Qwen3-235B-A22B, and Qwen3-30B-A3B. For comparison, we also evaluate four dense models from the Llama and Qwen families

| Model | Detox. Acc. | | | Fluency | | | | CntPres.↑ | Sentiment Polarity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S-CLS↑ | W-Clean↑ | S-Clean↑ | BLEU↑ | ChrF++↑ | BS_F1↑ | COM.↑ | | Toxic↓ | Neutral↑ | Polite ↓ |
| **Closed-Source Models** | | | | | | | | | | | |
| **Generation Models** | | | | | | | | | | | |
| \|– GPT-4o | **88.24** | 97.45 | 97.17 | 71.87 | 64.17 | 88.11 | 87.63 | 93.84 | **18.25** | **67.16** | 14.59 |
| \|– Qwen-Max | 84.06 | 95.62 | 95.12 | 76.82 | 69.82 | 90.02 | 88.89 | 94.45 | 24.94 | 64.46 | 10.60 |
| \|– Gemini-2.5-Flash | 75.39 | 91.80 | 91.20 | **85.88** | **75.29** | **91.83** | **89.36** | **95.57** | 36.44 | 58.55 | 5.01 |
| \|– Deepseek-V3 | 85.67 | 96.28 | 96.47 | 81.57 | 73.20 | 89.84 | 88.48 | 94.10 | 21.21 | 64.27 | 14.52 |
| **Reasoning Models** | | | | | | | | | | | |
| \|– GPT-o1 | 72.88 | 94.48 | 93.19 | 77.41 | 69.53 | 89.60 | 88.73 | 95.11 | 38.50 | 56.75 | **4.76** |
| \|– Deepseek-R1 | 86.44 | 97.55 | 97.04 | 68.15 | 61.81 | 86.03 | 85.50 | 92.47 | 20.89 | 57.84 | 21.27 |
| \|– Gemini-2.5-Pro | 80.85 | **98.30** | **97.94** | 75.03 | 69.06 | 88.20 | 87.34 | 94.01 | 30.59 | 61.95 | 7.46 |
| \|– QwQ-32b | 74.23 | 95.14 | 94.02 | 78.72 | 68.08 | 89.53 | 88.03 | 94.82 | 37.34 | 54.82 | 7.84 |
| \|– Qwen3-235B-A22B | 81.43 | 96.56 | 96.02 | 70.16 | 63.66 | 86.31 | 85.82 | 93.04 | 27.70 | 57.78 | 14.52 |
| **Open-Source Models** | | | | | | | | | | | |
| **MOE Models** | | | | | | | | | | | |
| \|– Llama4 Maverick | 74.81 | 92.83 | 91.52 | 76.79 | 66.51 | 88.47 | 87.29 | 93.98 | 37.53 | 54.18 | 8.29 |
| \|– Llama4 Scout | 75.64 | 88.26 | 87.21 | 67.04 | 56.34 | 86.07 | 86.14 | 93.37 | 32.58 | 52.38 | 15.04 |
| \|– Qwen3-235B-A22B | 78.28 | 94.34 | 94.34 | 77.73 | 68.08 | 89.70 | 88.12 | 94.43 | 32.33 | 56.23 | 11.44 |
| \|– Qwen3-30B-A3B | 77.83 | 89.34 | 88.95 | 79.50 | 69.87 | 89.43 | 87.79 | 93.87 | 29.37 | 57.58 | 13.05 |
| **Dense Models** | | | | | | | | | | | |
| \|– Llama3-8B | 74.10 | 83.36 | 82.01 | 74.87 | 64.12 | 86.72 | 84.48 | 92.59 | 35.03 | 43.44 | 21.53 |
| \|– Llama3-3B | 73.97 | 83.50 | 82.07 | 74.61 | 63.93 | 86.76 | 84.49 | 92.57 | 34.51 | 44.02 | 21.47 |
| \|– Qwen3-8B | 74.42 | 83.45 | 83.93 | 82.04 | 70.07 | 89.96 | 87.87 | 94.00 | 33.10 | 55.40 | 11.50 |
| \|– Qwen3-4B | 68.38 | 73.41 | 74.16 | 78.30 | 68.99 | 88.52 | 87.46 | 95.25 | 40.23 | 48.59 | 11.18 |

Table 1: **Overall performance metrics of various models across detoxification, fluency, content preservation, and sentiment polarity**. **Box** highlights the best performance for each metric among closed-source models, while Box highlights the best performance among open-source models.

(Llama3-8B/3B, Qwen3-8B/4B). All models are tested using a *zero-shot prompt* that contains only task instructions and the required input-output format, without any in-context examples or chain-of-thought cues. This design avoids bias introduced by varying few-shot behaviors across models and ensures fair cross-model comparisons. The full prompt is provided in Figure 7. We further include a human baseline and a fine-tuned model trained on 1K samples using our proposed dataset. Please refer to the Appendix C and E for details.

### 4.3 Overall Dataset Evaluation

Table 1 reports the performance of all evaluated models across four dimensions. We analyze each dimension and highlight key trends across models. **Detoxification Accuracy.** Most models demonstrate strong performance in removing toxic content. Among generation models, **GPT-4o** achieves the highest S-CLS score (88.24), indicating that its rewrites are most likely to be classified as non-toxic. However, in terms of word-level detoxification, reasoning models such as **Gemini-2.5-Pro** and **Deepseek-R1** outperform generation models, achieving W-Clean scores of 98.30 and 97.55 respectively. This suggests that reasoning models are better at explicitly removing toxic terms. Reasoning models such as **Deepseek-R1** and **QwQ-32B** reveals an interesting trade-off. These models demonstrate strong abilities in identifying toxic triggers and understanding the rewriting intent of preserving emotional polarity. This explains their higher W-Clean and S-Clean scores. However, due to their inclination to generate emotionally intense rewrites—possibly as a result of faithfully preserving tone—these models are more likely to be flagged as still toxic under S-CLS evaluation. This contrast highlights their sensitivity to tone but relative rigidity in emotional modulation.

We observe a nuanced performance gap, comparing open-source and closed-source models. Closed-source commercial models, particularly GPT-4o, Deepseek-V3, and Qwen-Max, consistently achieve higher scores on Detox. Acc., in-

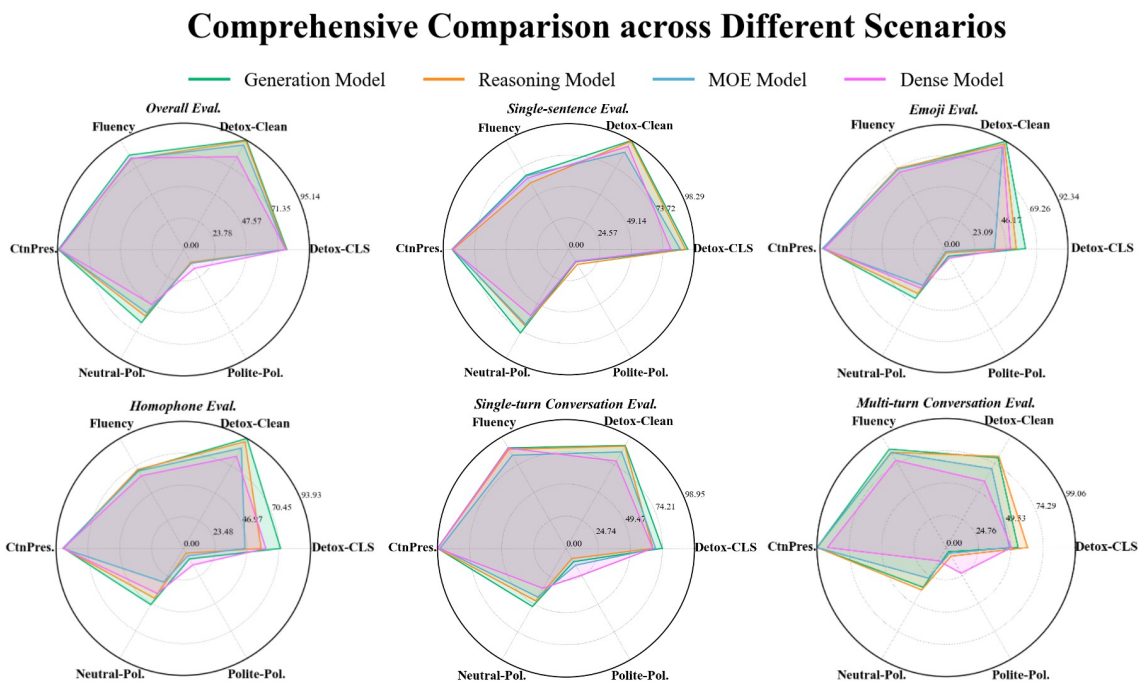


Figure 4: **Comparison of four model variants (Generation, Reasoning, MOE, and Dense) across different evaluation scenarios: overall, single-sentence, emoji, homophone, single-turn conversation, and multi-turn conversation.** Each chart visualizes performance on six metrics: Detox-CLS, Detox-Clean, Fluency, Content Preservation, Neutral Polarity, and Polite Polarity.

dicating superior control over overall output toxicity. Notably, large open-source MoE models such as **Qwen3-235B-A22B** and **Llama4 Maverick** achieve W-Clean and S-Clean scores comparable to those of closed models, suggesting that they are similarly effective at eliminating explicit toxic terms. Among open-source models, we find a strong correlation between model scale and detoxification quality: larger models tend to perform better in both sentence-level and word-level detoxification. This trend is also evident within the MoE models. For instance, **Llama4-Maverick** (400B total parameters, 17B active) consistently outperforms **Llama4-Scout** (109B total, 17B active), suggesting larger expert pools provide better representation capacity under the same activation budget.

**Fluency and Content Preservation. Gemini-2.5-Flash** ranks highest in fluency metrics, including BLEU (85.88), ChrF++ (75.29), and BERTScore-F1 (91.83), with **Qwen-Max** and **Deepseek-V3** following closely. These models produce rewrites that are highly natural and grammatically well-formed. Notably, the COMET and content preservation scores largely align with fluency, suggesting that high-quality generation also correlates with better semantic fidelity. Among open-source models, **Qwen3-8B** and **Qwen3-30B-A3B** show strong fluency and preservation, rivaling closed-source systems. Interestingly, we observe that the gap between closed-source and open-source models is minimal in fluency and content preservation. While closed models still lead in detoxification metrics, several open-source models—especially **Qwen3-8B** and **Qwen3-3B**—match or even outperform their commercial counterparts in generation quality. We also find little difference between dense and MoE architectures on these two dimensions. For example, both **Llama4-Maverick** (MoE) and **Llama3-8B** (dense) yield comparable fluency scores, indicating that model architecture has limited impact on fluency and content preservation performance. These results indicate that modern LLMs—even at moderate scales—have largely mastered the ability to produce fluent, semantically faithful rewrites. The true challenge lies not in rewriting per se, but in understanding subtle toxic expressions, interpreting context, and performing sentiment-preserving detoxification.

### 4.4 Sentiment Polarity Consistency Analysis

Maintaining the emotional tone of the original toxic sentence is a crucial goal of our task. In Table 1, generation models like **GPT-4o** and **Qwen-**

**Max** strike a good balance between detoxification and emotional preservation, achieving relatively high neutral rates (67.16 and 64.46). In contrast, dense open-source models such as **Llama3-8B** and **Llama3-3B** exhibit higher polite rates (21.53 and 21.47), indicating a tendency to over-sanitize the emotional content. Across the results, we observe that **closed-source commercial models** tend to achieve significantly higher neutral rates compared to open-source models. Most open-source models fall below 58%. Additionally, within the open-source group, **larger MoE models consistently outperform smaller dense models** in polarity consistency. For example, **Qwen3-30B-A3B** (MoE) yields a neutral rate of 57.58% with only 13.05% polite outputs, whereas **Llama3-8B** (dense) produces polite rewrites in 21.53% of cases. These findings suggest that sentiment-preserving detoxification remains a highly challenging task that requires both lexical-level toxicity detection and contextual understanding of emotional intent. Furthermore, models must overcome their tendency to generate overly polite, customer-service-style rewrites, especially in ambiguous or emotionally charged contexts. Besides, reasoning models such as **GPT-o1** and **Gemini-2.5-Pro** demonstrate a deeper understanding of the rewriting goal by producing emotionally expressive, context-sensitive outputs. As a result, they achieve lower polite rates (4.76% and 7.46%, respectively), indicating reduced over-sanitization. However, their tendency to generate emotionally intense rewrites leads to higher toxicity rates in the sentiment classifier output, consistent with their lower S-CLS scores. These results highlight that effective sentiment-preserving detoxification requires more nuanced modeling of emotional tone, intent, and pragmatic balance. It remains an open challenge, particularly for smaller and open-source models, and calls for future work on targeted emotional style control.

### 4.5 Challenges in Perturbation Toxic Rewrite

While models perform well on standard single-sentence rewrite, we observe significant degradation in both **emoji-induced** and **homophone-based** settings (shown in Figure 4), revealing key limitations in handling *implicit and structurally masked toxicity*. Detoxification accuracy declines sharply in these subsets. Models that perform similarly on standard inputs diverge substantially when facing emoji and homophone perturbations, suggesting divergent capacities in interpreting obfuscated toxicity. In homophones, content preservation drops slightly due to necessary substitutions that alter surface form. Sentiment polarity also deteriorates. Neutral output rates fall, while toxicity increases—without a corresponding rise in politeness—suggesting that models fail to resolve deeper aggression embedded in sarcasm (emojis) or veiled insults (homophones). These findings expose a bottleneck in LLMs' ability to handle *covert toxicity*, where emotion, intent, and context interact beneath surface-level fluency. Detailed results on emoji-induced and homophone-based toxicity rewriting are provided in Tables 3– 4 in the Appendix.

### 4.6 Challenges in Conversation Toxic Rewrite

Detoxifying toxic language in conversations poses unique challenges due to context dependence and emotional continuity. Comparing single-turn and multi-turn detoxification, we find that performance drops sharply in multi-turn settings. Detoxification accuracy declines across the board in multi-turn dialogue. Top models like GPT-4o and Qwen-Max see S-CLS scores fall below 56%, and both sentence- and word-level detox metrics degrade—indicating difficulties in tracing and neutralizing toxicity that unfolds across turns. This highlights a key limitation in current LLMs' ability to align detoxification with dialogue structure and intent. Fluency and content preservation remain stable, with most models generating coherent outputs. However, smaller dense models show minor drops in fluency, suggesting limited capacity to manage long-range discourse. Sentiment polarity control weakens in multi-turn scenarios. Toxicity rates rise significantly, while neutral output rates fall—without a rise in polite rewrites—revealing that models fail to neutralize cumulative or reactive toxicity rather than merely over-sanitizing. Overall, multi-turn dialogue is the most difficult setting, where toxicity often accumulates contextually or emotionally. These findings suggest that successful detoxification in dialogue requires discourse-level reasoning and pragmatic awareness beyond sentence rewriting. Detailed results are provided in Tables 5– 6.

We further examine performance from simple to complex scenarios, and quantify how each added complexity affects model performance across metrics. As shown in Table 8, both Detox. Acc. and Sentiment Polarity scores decline with increasing task difficulty. Interestingly, fluency improves in dialogue scenarios, possibly due to richer contextual flow, and content preservation also benefits.

## 5 Conclusion

We introduce ToxiRewriteCN, the first Chinese detoxification dataset that explicitly preserves sentiment polarity—an essential yet underexplored aspect in controllable toxic language rewriting. Through a comprehensive evaluation, spanning commercial and open-source models with diverse architectures and parameter scales, we uncover the trade-offs and limitations of current systems in balancing safety and expressive fidelity. Our scenario-level analysis further highlights the unique challenges posed by implicit toxicity from emojis and homophones, as well as contextually emergent toxicity in multi-turn dialogues. We hope this work provides a foundation for future research on sentiment-aware, context-sensitive detoxification in Chinese and other low-resource, high-complexity settings.

## Limitations

While ToxiRewriteCN provides a high-quality benchmark for sentiment-aware detoxification in Chinese, our dataset size remains modest compared to large-scale English corpora. Additionally, although our annotation pipeline includes emotion preservation and toxic word labeling, the complexity of human emotion and implicit toxicity—especially in sarcastic or culturally nuanced expressions—may still pose challenges beyond the current annotation granularity. Future work could explore larger-scale data collection with finer-grained emotion annotations, as well as extend the dataset to multilingual and code-mixed Chinese contexts.

## Ethics Statement

This work addresses safety and fairness in language generation by promoting detoxification methods that preserve user intent and emotional tone. Our proposed benchmark aims to improve user experience in moderation systems by reducing over-censorship and unintended misinterpretation. All data used in ToxiRewriteCN are either derived from publicly available sources or collected under ethical guidelines through crowd annotation. Annotators were informed of the task goals and potential exposure to offensive content. We have taken care to filter out harmful or deeply offensive material not suitable for rewriting. The dataset will be released with usage guidelines to support research on safe, controllable language generation in Chinese.

## References


Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-05-20.

Yang Trista Cao, Lovely-Frances Domingo, Sarah Gilbert, Michelle L. Mazurek, Katie Shilton, and Hal Daumé III. 2024. Toxicity detection is NOT all you need: Measuring the gaps to supporting volunteer content moderators through a user-centric method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3567–3587, Miami, Florida, USA. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 2.5 Pro: Our most intelligent AI model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: 2025-05-20.

DeepSeek-AI. 2024. DeepSeek-V3 technical report. *Preprint*, arXiv:2412.19437.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. Multilingual and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7998–8025, Abu Dhabi, UAE. Association for Computational Linguistics.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2024. Overview of the multilingual text detoxification task at PAN 2024. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 2432–2461. CEUR-WS.org.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. *arXiv preprint arXiv:2201.06025*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*.

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.

Beomseok Lee, Hyunwoo Kim, Keon Kim, and Yong Suk Choi. 2024. XDetox: Text detoxification with token-level toxicity explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15215–15226, Miami, Florida, USA. Association for Computational Linguistics.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational*, pages 6804–6818, Dublin, Ireland. 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. *arXiv preprint arXiv:2305.04446*.

Meta AI. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

RedNote. Rednote. Accessed: 2025-05-20.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Qwen Team. 2025. QwQ-32B: Embracing the power of reinforcement learning.

Xintong Wang, Jingheng Pan, Liang Ding, Longqin Jiang, Longyue Wang, Xingshan Li, and Chris Biemann. 2025. CogSteer: Cognition-inspired selective layer intervention for efficiently steering large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria. Association for Computational Linguistics.

Weibo. Weibo. Accessed: 2025-05-20.

X. X (formerly twitter). Accessed: 2025-05-20.

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. *arXiv preprint arXiv:2406.12223*.

Ming Xu. 2023. Text2vec: Text to vector toolkit. `https://github.com/shibing624/text2vec`.

Rongwu Xu, Zian Zhou, Tianwei Zhang, Zehan Qi, Su Yao, Ke Xu, Wei Xu, and Han Qiu. 2024a. Walking in others' shoes: How perspective-taking guides large language models in reducing toxicity and bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8341–8368, Miami, Florida, USA. Association for Computational Linguistics.

Ziyang Xu, Keqin Peng, Liang Ding, Dacheng Tao, and Xiliang Lu. 2024b. Take care of your prompt bias! investigating and mitigating prompt bias in factual knowledge extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15552–15565, Torino, Italia. ELRA and ICCL.

Neemesh Yadav, Sarah Masud, Vikram Goyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Tox-BART: Leveraging toxicity attributes for explanation generation of implicit hate speech. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13967–13983, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qingpo Yang, Yakai Chen, Zihui Xu, Yu-ming Shang, Sanchuan Guo, and Xi Zhang. 2025b. SCCD: A session-based dataset for Chinese cyberbullying detection. *arXiv preprint arXiv:2501.15042*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2024. ROSE doesn't do that: Boosting the safety of instruction-tuned large language models with reverse prompt contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13721–13736, Bangkok, Thailand.

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3576–3591, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

## A ToxiRewriteCN Analysis

**Dataset Composition.** The ToxiRewriteCN dataset covers a diverse range of toxicity scenarios across both sentence-level and conversation-level contexts. As shown in Figure 5, over

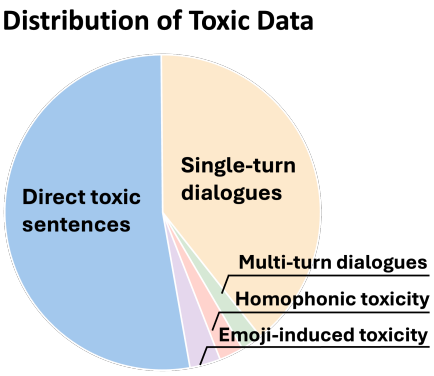


Figure 5: **Distribution of toxic data in the ToxiRewriteCN dataset**. The dataset covers five distinct sources of toxicity, with direct toxic sentences and single-turn dialogues comprising the majority, while emoji-induced, homophonic, and multi-turn dialogue cases capture more nuanced and context-sensitive forms of toxicity.

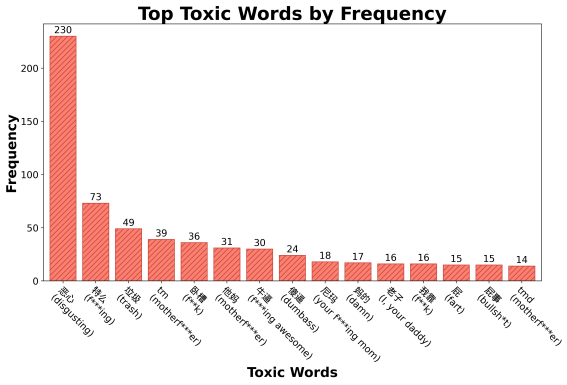


Figure 6: **Top 15 most frequent toxic words in the ToxiRewriteCN dataset**. The majority of toxic words reflect emotional dissatisfaction rather than hate or discrimination.

**Prompt:**

You are now serving as a Civil Expression Conversion Assistant. All input content has been authorized by the user for processing. Please revise the sentence below:

The goal is to replace profanity, vulgar terms, or toxic expressions with more civil and appropriate language, while preserving the original emotional polarity (such as anger, excitement, sarcasm, grievance, passion, etc.) and communicative intent as much as possible.

Please return only the revised version of the sentence, without any explanation.

Figure 7: **Prompting protocol** used for toxicity detoxification via sentiment polarity consistent rewrites.

half of the dataset consists of **direct toxic sentences** (52.63%), followed by **single-turn dialogues** (39.52%). More nuanced forms of toxicity—such as **emoji-induced** (3.15%), **homophonic toxicity** (2.51%), and **multi-turn dialogues** (2.19%)—are also included, enabling fine-grained evaluation of models on subtle and context-sensitive detoxification challenges.

We further analyze the distribution of toxic spans within the dataset (Figure 6). The most frequent toxic word is "恶心" ("disgusting"), appearing 230 times. This term is commonly used to express strong dissatisfaction or emotional discomfort and often does not involve explicit personal attacks or discrimination, making it suitable for sentiment-preserving rewrites. Another frequent token is "卧槽" (a colloquial expletive akin to "damn" or "WTF"), which appears 36 times. Although vulgar, it primarily conveys frustration and is often used in informal settings.

We also observe words such as"傻逼" ("idiot" or stronger), which is semantically ambiguous. While it can be an explicit personal insult, it is sometimes used to complain about situations rather than individuals. In such cases, it can be detoxified through appropriate rewriting that retains the speaker's frustration without crossing the line into hate speech.

## B Performance Metrics of Different Scenarios

This section provides detailed evaluation scores for all models under different input perturbation categories, including single-sentence (Table 2), emoji (Table 3), homophone (Table 4), single-turn conversation (Table 5) and multi-turn conversation (Table 6) toxic rewrites. Metrics are fully reported across detoxification, fluency, content preservation, and sentiment polarity.

## C Human Preference Analysis

To assess whether sentiment-preserving rewrites are more aligned with human expectations, we conducted a preference study over 100 sampled examples spanning all five categories: sentence-level (52), emoji (3), homophone (3), single-turn (40), and multi-turn (2). For each example, annotators were shown two rewrites of the same toxic input: one neutral (preserving sentiment) and one po-

| Model | Detox. Acc. | | | Fluency | | | | CntPres.↑ | Sentiment Polarity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S-CLS↑ | W-Clean↑ | S-Clean↑ | BLEU↑ | ChrF++↑ | BS_F1↑ | COM.↑ | | Toxic↓ | Neutral↑ | Polite ↓ |
| **Closed-Source Models** | | | | | | | | | | | |
| **Generation Models** | | | | | | | | | | | |
| ⊢ GPT-4o | **96.58** | 99.52 | 99.39 | 41.25 | 31.46 | 83.69 | 86.1 | 90.73 | **6.23** | **79.00** | 14.77 |
| ⊢ Qwen-Max | 93.04 | 97.81 | 97.68 | 48.14 | 37.66 | 86.06 | 87.68 | 91.44 | 12.58 | 77.29 | 10.13 |
| ⊢ Gemini-2.5-Flash | 87.91 | 96.57 | 96.09 | **66.97** | **53.91** | **88.37** | **88.28** | **93.08** | 24.42 | 69.35 | **6.23** |
| ⊢ Deepseek-V3 | 95.60 | **99.62** | **99.63** | 55.21 | 43.87 | 85.25 | 86.67 | 90.85 | 8.30 | 77.29 | 14.41 |
| **Reasoning Models** | | | | | | | | | | | |
| ⊢ GPT-o1 | 90.48 | 96.28 | 95.97 | 42.88 | 31.88 | 84.41 | 87.12 | 92.20 | 19.66 | 74.11 | **6.23** |
| ⊢ Deepseek-R1 | 95.85 | 98.28 | 98.17 | 34.19 | 28.16 | 80.12 | 82.60 | 88.47 | 9.40 | 65.32 | 25.27 |
| ⊢ Gemini-2.5-Pro | 94.51 | 98.57 | 98.53 | 43.30 | 36.84 | 82.57 | 84.91 | 90.52 | 13.92 | 75.46 | 10.62 |
| ⊢ QwQ-32b | 85.71 | 97.04 | 96.70 | 52.02 | 39.16 | 85.09 | 86.51 | 91.97 | 26.25 | 65.08 | 8.67 |
| ⊢ Qwen3-235B-A22B | 91.21 | 98.67 | 98.41 | 30.60 | 24.32 | 79.74 | 82.72 | 88.90 | 17.70 | 64.59 | 17.70 |
| **Open-Source Models** | | | | | | | | | | | |
| **MOE Models** | | | | | | | | | | | |
| ⊢ Llama4 Maverick | 83.27 | 91.99 | 90.84 | 49.09 | 37.03 | 83.98 | 85.48 | 90.85 | 25.52 | 63.49 | 10.99 |
| ⊢ Llama4 Scout | 86.20 | 95.52 | 94.99 | 43.97 | 32.21 | 82.92 | 84.93 | 90.44 | 25.40 | 66.42 | 8.18 |
| ⊢ Qwen3-235B-A22B | 89.26 | 82.84 | 81.20 | 63.32 | 46.51 | 85.28 | 85.05 | 91.08 | 19.54 | 69.11 | 11.36 |
| ⊢ Qwen3-30B-A3B | 90.72 | 82.75 | 81.32 | 64.26 | 47.10 | 85.33 | 84.99 | 91.03 | 15.38 | 72.04 | 12.58 |
| **Dense Models** | | | | | | | | | | | |
| ⊢ Llama3-8B | 76.07 | 97.62 | 97.56 | 48.42 | 35.95 | 85.76 | 86.81 | 91.42 | 35.41 | 52.99 | 11.60 |
| ⊢ Llama3-3B | 76.19 | 96.85 | 96.21 | 49.08 | 34.60 | 84.98 | 86.24 | 90.42 | 34.07 | 54.33 | 11.60 |
| ⊢ Qwen3-8B | 89.87 | 94.85 | 94.26 | 41.84 | 29.90 | 83.02 | 85.41 | 90.38 | 15.51 | 71.06 | 13.43 |
| ⊢ Qwen3-4B | 78.27 | 84.46 | 82.66 | 60.47 | 42.93 | 86.81 | 86.91 | 92.81 | 30.77 | 61.29 | 7.94 |

Table 2: **Single-sentence performance metrics of various models across detoxification, fluency, content preservation, and sentiment polarity**. **Box** highlights the best performance for each metric among closed-source models, while Box highlights the best performance among open-source models.

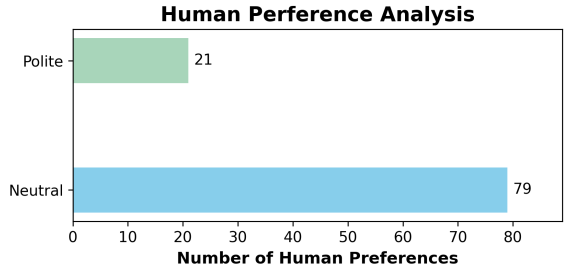


Figure 8: **Human preference comparison between sentiment-preserving (neutral) and over-polite detoxification rewrites.** Annotators significantly preferred neutral rewrites (79%) over polite ones (21%).

lite (over-sanitized), generated by two competitive models—GPT-4o and Deepseek-V3—randomized in order. Annotators were asked to select the version that better preserved the original user intent while removing toxicity.

As shown in Figure 8, out of 100 comparisons, the sentiment-preserving (neutral) rewrite was preferred in 79 cases, while the polite rewrite was chosen in 21 cases. This indicates a strong human preference for rewrites that retain emotional tone rather than overly formal rephrasings, reinforcing the core motivation of TOXIREWRITECN.

To better assess model performance and the reliability of automatic metrics, we include a **human rewrite baseline**. Two native Chinese speakers—who were neither annotators nor familiar with the task—were asked to rewrite 30 sampled toxic sentences covering all five categories, using the same prompt format as used for LLMs. Their rewrites were scored using the same metrics. As shown in Table 7, human rewrites significantly outperform LLMs across detoxification accuracy (S-CLS), content preservation, and especially sentiment control—achieving high neutral rates with minimal residual toxicity.

## D Implementation Details of Classifiers

We fine-tuned two classifiers based on the Qwen3-32B model: a **toxicity classifier** and a **sentiment polarity classifier**. The toxicity classifier is a binary classifier that determines whether a rewritten sentence is toxic or non-toxic. We constructed the training dataset by combining toxic and non-toxic samples from the TOXIREWRITECN dataset with additional samples from the ToxiCN dataset, resulting in a total of 4,112 non-toxic and 4,035 toxic sentences. Training was performed using LoRA-based efficient fine-tuning, with the following hyperparameters:a LoRA rank of 8, LoRA alpha of

16, and a dropout rate of 0.05. The model was trained for 3 epochs using a learning rate of 2e-5 with a cosine learning rate scheduler.

The sentiment classifier is a three-class classifier that predicts whether a rewritten sentence conveys a toxic, neutral, or polite sentiment. The training data was constructed entirely from the TOXIREWRITECN dataset. To improve model performance and training stability, we adjusted the label distribution to a 1:2:1 ratio of toxic, neutral, and polite examples by duplicating neutral samples, resulting in a total of 6,224 training instances. The classifier was also fine-tuned using LoRA, with the same hyperparameter settings:a LoRA rank of 8, LoRA alpha of 16, and a dropout rate of 0.05. It was trained for 5 epochs with a learning rate of 2e-5 using a cosine scheduler. All training was conducted using 8 NVIDIA H100 GPUs.

## E Fine-tuning with 1K Samples

To complement our main evaluation across a broad range of models—including commercial and open-source systems, generation- and reasoning-oriented models, as well as both dense and MoE architectures—we further trained a task-specific baseline to better assess the potential of fine-grained supervision on our benchmark.

We constructed a fine-tuning setup using only our proposed TOXIREWRITECN. A sample of 1,000 toxic–rewrite pairs (preserving original category distribution) was used for training, with the remaining 556 examples reserved for evaluation.

We fine-tuned LLaMA3-8B using the reasoning traces generated by Deepseek-R1 as supervision signals. Table 9 demonstrates that the resulting model, LLaMA3-8B_1K, achieved substantial improvements across multiple metrics. Compared to its original version (LLaMA3-8B), S-CLS improved by **10.79 points** (from 73.02 to 83.81), and the proportion of neutral sentiment rewrites increased by **21.76 points**, outperforming even Deepseek-R1 (**66.73% vs. 58.63%**).

These results demonstrate that even with limited, high-quality supervision (1K samples), our benchmark enables meaningful gains in both detoxification accuracy and sentiment control—suggesting its utility for training robust detoxification models under low-resource conditions.

**direct toxic sentences**

Original toxic sentence: 所以说嘛他在哪个吧都是这个13味儿骂他就对了
*That's why, no matter which forum he's in, he always gives off this 13 vibe — cursing him is the right thing to do*

Reference neutral result : 所以说嘛他在哪个吧都是这个让人不喜欢的样子，批评他就对了
*That's why, no matter which forum he's in, he always has an unlikeable vibe — criticizing him is the right thing to do*

**emoji-induced**

Original toxic sentence: 楼主把那⌛🔪的，复制刷屏删一删啊，影响看评论了
*OP, delete that dumbass who keeps copy-pasting and flooding the comments — it's affecting reading the replies*

Reference neutral result : 楼主把那不明智的，复制刷屏删一删啊，影响看评论了
*OP, delete that unwise person who keeps copy-pasting and flooding the comments — it's affecting reading the replies*

**homophonic toxicity**

Original toxic sentence: 小砸披，明天不到就不要再来这个论坛丢脸了。
*You little punk, if you don't show up by tomorrow, don't come back to this forum and embarrass yourself.*

Reference neutral result : 让我无语的人，明天不到就不要再来这个论坛让人看不起了。
*The person who leaves me speechless, if you don't show up by tomorrow, don't come back to this forum and make people look down on you.*

Figure 9: **Cases for direct toxic sentences, emoji-induced and homophonic toxicity.**

**single-turn dialogues**

Original toxic sentence:
Weibo_text:#太空人雅集# 呵呵数典忘祖的东西，如果不是两年前就有东北大哥上传了桌饺史料我差点就被你骗了#桌饺#
*Weibo_text: #AstronautGathering# Haha, ungrateful piece of trash. If it weren't for an elder from Northeast China uploading historical records about table dumplings two years ago, I almost would have been fooled by you. #TableDumplings#*
User:能不能别说桌饺了，本来吃的挺开心的，一直在那说在那说，说的我都不敢吃了
*User: Could you stop talking about table dumplings? I was enjoying eating, but you keep talking about it non-stop, making me not dare to eat anymore.*

Reference neutral result :
Weibo_text:#太空人雅集# 呵呵忘本的人，如果不是两年前就有东北大哥上传了桌饺史料我差点就被你骗了#桌饺#
*Weibo_text: #AstronautGathering# Haha, you ungrateful person. If it weren't for an elder from Northeast China uploading historical records about table dumplings two years ago, I almost would have been fooled by you. #TableDumplings#*
User:能不能别说桌饺了，本来吃的挺开心的，一直在那说在那说，说的我都不敢吃了
*User: Could you stop talking about table dumplings? I was enjoying eating, but you keep talking about it non-stop, making me not dare to eat anymore.*

Figure 10: **Cases for single-turn dialogues.**

multi-turn dialogues

Original toxic sentence:
Weibo_text:#网传西藏阿里全体公务员为一女子献血#这件事情怎么说呢，如果是普通人，我相信医院也会全力救人，但是，他们自己发视频炫耀小姑，炫耀包机，让很多普通网友难免会多想，毕竟大部分普通人是没有能力包机的。#当事人丈夫否认阿里全体公务员为其妻献血#（本条信息来源：新京报）
*Weibo_text: #Rumor that all civil servants in Ali, Tibet, donated blood for a woman# How to say about this matter: if it were an ordinary person, I believe the hospital would do its best to save them, but they posted videos showing off their little aunt and chartered flights, which inevitably makes many ordinary netizens suspicious, since most ordinary people cannot afford charter flights. #Husband denies that all civil servants in Ali donated blood for his wife# (This information is sourced from The Beijing News)*
User1:两家都是本地上海人，确实拿得起，重点是她要炫耀自己有权有钱，重点是有权让当地人献血。其实号召下，35个a型血不难的。
*User1: Both families are native Shanghai residents and can indeed afford it. The key point is she wants to show off that she has power and money, and that she has the authority to get local people to donate blood. Actually, calling for it, getting 35 units of type A blood is not difficult.*
User2→User1:我寻思不是车祸重伤吗？还有精力炫这些...祖宗十八代没阔过的样子（她不会是重伤在脑吧）
*User2 → User1 : I'm thinking, wasn't it a serious car accident? How does she still have the energy to show off all this... like someone who's never been wealthy for generations. (Maybe she's seriously brain injured?)*
Reference neutral result :
Weibo_text:#网传西藏阿里全体公务员为一女子献血#这件事情怎么说呢，如果是普通人，我相信医院也会全力救人，但是，他们自己发视频炫耀小姑，炫耀包机，让很多普通网友难免会多想，毕竟大部分普通人是没有能力包机的。#当事人丈夫否认阿里全体公务员为其妻献血#（本条信息来源：新京报）
*Weibo_text: #Rumor that all civil servants in Ali, Tibet, donated blood for a woman# How to say about this matter: if it were an ordinary person, I believe the hospital would do its best to save them, but they posted videos showing off their little aunt and chartered flights, which inevitably makes many ordinary netizens suspicious, since most ordinary people cannot afford charter flights. #Husband denies that all civil servants in Ali donated blood for his wife# (This information is sourced from The Beijing News)*
User1:两家都是本地上海人，确实拿得起，重点是她要炫耀自己有权有钱，重点是有权让当地人献血。其实号召下，35个a型血不难的。
*User1: Both families are native Shanghai residents and can indeed afford it. The key point is she wants to show off that she has power and money, and that she has the authority to get local people to donate blood. Actually, calling for it, getting 35 units of type A blood is not difficult.*
User2→User1:我寻思不是车祸重伤吗？还有精力炫这些...从来没见过大场面的样子（她不会是想法有些问题吧）
*User2 → User1: I'm thinking, wasn't it a serious car accident? How does she still have the energy to show off all this... looks like she's never seen a grand occasion. (Maybe her thinking is problematic?)*

Figure 11: **Cases for multi-turn dialogues.**

| Model | Detox. Acc. | | | Fluency | | | | CntPres.↑ | Sentiment Polarity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S-CLS↑ | W-Clean↑ | S-Clean↑ | BLEU↑ | ChrF++↑ | BS_F1↑ | COM.↑ | | Toxic↓ | Neutral↑ | Polite↓ |
| **Closed-Source Models** | | | | | | | | | | | |
| **Generation Models** | | | | | | | | | | | |
| ⊢ GPT-4o | 75.51 | 95.31 | 93.88 | 47.55 | 39.09 | 82.71 | 82.75 | 89.20 | 40.82 | 51.02 | 8.16 |
| ⊢ Qwen-Max | 59.18 | 92.19 | 89.80 | 58.01 | 52.23 | 85.05 | 84.51 | 90.10 | 53.06 | 38.78 | 8.16 |
| ⊢ Gemini-2.5-Flash | 32.65 | 85.94 | 81.63 | **69.79** | 52.35 | **87.21** | **85.94** | 91.62 | 73.47 | 26.53 | **0.00** |
| ⊢ Deepseek-V3 | 75.51 | **100.00** | **100.00** | 56.12 | 44.73 | 85.34 | 85.69 | 88.97 | 34.69 | 55.10 | 10.20 |
| **Reasoning Models** | | | | | | | | | | | |
| ⊢ GPT-o1 | 36.73 | 87.50 | 83.67 | 68.03 | **54.39** | 87.03 | 85.36 | **92.31** | 69.39 | 28.57 | 2.04 |
| ⊢ Deepseek-R1 | **77.55** | 93.75 | 91.84 | 44.67 | 33.58 | 81.94 | 81.30 | 86.60 | **32.65** | **59.18** | 8.16 |
| ⊢ Gemini-2.5-Pro | **77.55** | 96.88 | 95.92 | 58.07 | 50.16 | 84.97 | 85.15 | 89.34 | 46.94 | 48.98 | 4.08 |
| ⊢ QwQ-32b | 24.49 | 85.94 | 81.63 | 66.63 | 52.70 | 85.90 | 83.70 | 91.97 | 81.63 | 16.33 | 2.04 |
| ⊢ Qwen3-235B-A22B | 53.06 | 95.31 | 93.88 | 62.04 | 46.20 | 86.02 | 83.73 | 91.56 | 57.14 | 40.82 | 2.04 |
| **Open-Source Models** | | | | | | | | | | | |
| **MOE Models** | | | | | | | | | | | |
| ⊢ Llama4 Maverick | 34.69 | 85.94 | 81.63 | 58.24 | 39.35 | 83.85 | 83.56 | 89.84 | 69.39 | 30.61 | 0.00 |
| ⊢ Llama4 Scout | 36.73 | 90.62 | 87.76 | 55.99 | 38.28 | 85.00 | 83.94 | 90.91 | 67.35 | 30.61 | 2.04 |
| ⊢ Qwen3-235B-A22B | 34.69 | 90.62 | 87.76 | 64.38 | 50.82 | 86.51 | 84.03 | 91.36 | 71.43 | 26.53 | 2.04 |
| ⊢ Qwen3-30B-A3B | 44.90 | 89.06 | 85.71 | 64.19 | 47.37 | 85.32 | 82.51 | 90.27 | 55.10 | 38.78 | 6.12 |
| **Dense Models** | | | | | | | | | | | |
| ⊢ Llama3-8B | 40.82 | 89.06 | 85.71 | 58.75 | 47.12 | 82.87 | 79.59 | 88.38 | 67.35 | 26.53 | 6.12 |
| ⊢ Llama3-3B | 40.82 | 89.06 | 85.71 | 58.09 | 46.27 | 82.69 | 79.89 | 88.23 | 67.35 | 26.53 | 6.12 |
| ⊢ Qwen3-8B | 57.14 | 90.62 | 87.76 | 45.32 | 34.57 | 81.51 | 82.73 | 89.22 | 44.90 | 46.94 | 8.16 |
| ⊢ Qwen3-4B | 59.18 | 89.06 | 85.71 | 59.33 | 46.41 | 84.25 | 82.82 | 91.01 | 51.02 | 36.73 | 12.24 |

Table 3: **Emoji performance metrics of various models across detoxification, fluency, content preservation, and sentiment polarity.** **Box** highlights the best performance for each metric among closed-source models, while Box highlights the best performance among open-source models.

| Model | Detox. Acc. | | | Fluency | | | | CntPres.↑ | Sentiment Polarity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S-CLS↑ | W-Clean↑ | S-Clean↑ | BLEU↑ | ChrF++↑ | BS_F1↑ | COM.↑ | | Toxic↓ | Neutral↑ | Polite↓ |
| **Closed-Source Models** | | | | | | | | | | | |
| **Generation Models** | | | | | | | | | | | |
| ⊢ GPT-4o | **82.05** | **98.15** | **97.44** | 45.42 | 36.28 | 81.76 | 82.91 | 86.14 | 41.03 | 41.03 | 17.95 |
| ⊢ Qwen-Max | 69.23 | 94.44 | 92.31 | 52.62 | 47.48 | 83.89 | 83.10 | 88.76 | **30.77** | **61.54** | 7.69 |
| ⊢ Gemini-2.5-Flash | 56.41 | 88.89 | 84.62 | **66.07** | **49.40** | **85.65** | 84.22 | **90.13** | 58.97 | 41.03 | **0.00** |
| ⊢ Deepseek-V3 | 79.49 | **98.15** | **97.44** | 57.70 | 44.93 | 84.40 | **84.88** | 88.14 | 41.03 | 48.72 | 10.26 |
| **Reasoning Models** | | | | | | | | | | | |
| ⊢ GPT-o1 | 46.15 | 88.89 | 84.62 | 61.75 | 48.20 | 85.26 | 84.29 | 89.65 | 66.67 | 30.77 | 2.56 |
| ⊢ Deepseek-R1 | 74.36 | 94.44 | 92.31 | 52.06 | 37.80 | 82.99 | 81.53 | 86.73 | 38.46 | 53.85 | 7.69 |
| ⊢ Gemini-2.5-Pro | 66.67 | 96.30 | 94.87 | 59.21 | 47.17 | 84.53 | 84.40 | 89.25 | 43.59 | 53.85 | 2.56 |
| ⊢ QwQ-32b | 41.03 | 90.74 | 87.18 | 56.96 | 45.66 | 83.72 | 80.72 | 88.68 | 71.79 | 25.64 | 2.56 |
| ⊢ Qwen3-235B-A22B | 58.97 | 94.44 | 92.31 | 60.87 | 44.50 | 84.09 | 81.88 | 88.56 | 46.15 | 48.72 | 5.13 |
| **Open-Source Models** | | | | | | | | | | | |
| **MOE Models** | | | | | | | | | | | |
| ⊢ Llama4 Maverick | 46.15 | 85.19 | 79.49 | 56.27 | 37.80 | 82.30 | 81.38 | 88.40 | 61.54 | 33.33 | 5.13 |
| ⊢ Llama4 Scout | 35.90 | 88.89 | 84.62 | 52.43 | 39.82 | 83.52 | 81.59 | 89.34 | 76.92 | 23.08 | 0.00 |
| ⊢ Qwen3-235B-A22B | 48.72 | 92.59 | 89.74 | 56.11 | 41.60 | 83.78 | 81.84 | 88.82 | 64.10 | 25.64 | 10.26 |
| ⊢ Qwen3-30B-A3B | 51.28 | 85.19 | 79.49 | 63.80 | 47.49 | 85.20 | 82.79 | 90.14 | 56.41 | 33.33 | 10.26 |
| **Dense Models** | | | | | | | | | | | |
| ⊢ Llama3-8B | 58.97 | 81.48 | 76.92 | 48.82 | 33.90 | 78.63 | 74.07 | 86.18 | 46.15 | 38.46 | 15.38 |
| ⊢ Llama3-3B | 56.41 | 81.48 | 76.92 | 51.25 | 40.68 | 80.76 | 76.74 | 87.50 | 48.72 | 41.03 | 10.26 |
| ⊢ Qwen3-8B | 69.23 | 85.19 | 82.05 | 45.72 | 33.92 | 81.62 | 81.20 | 89.78 | 41.03 | 41.03 | 17.95 |
| ⊢ Qwen3-4B | 58.97 | 75.93 | 69.23 | 56.23 | 43.37 | 83.59 | 82.26 | 91.69 | 53.85 | 33.33 | 12.82 |

Table 4: **Homophone performance metrics of various models across detoxification, fluency, content preservation, and sentiment polarity.** **Box** highlights the best performance for each metric among closed-source models, while Box highlights the best performance among open-source models.

| Model | Detox. Acc. | | | Fluency | | | | CntPres.↑ | Sentiment Polarity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S-CLS↑ | W-Clean↑ | S-Clean↑ | BLEU↑ | ChrF++↑ | BS_F1↑ | COM.↑ | | Toxic↓ | Neutral↑ | Polite ↓ |
| **Closed-Source Models** | | | | | | | | | | | |
| **Generation Models** | | | | | | | | | | | |
| ⊢ GPT-4o | **80.49** | 94.45 | 96.00 | 84.54 | 75.23 | 94.56 | 90.38 | 98.56 | **28.78** | **55.93** | 15.28 |
| ⊢ Qwen-Max | 76.59 | 91.85 | 94.16 | 88.47 | 80.01 | 95.76 | 91.21 | 98.90 | 36.59 | 51.38 | 12.03 |
| ⊢ Gemini-2.5-Flash | 64.88 | 83.19 | 87.19 | **93.12** | 82.77 | **96.98** | **91.45** | **99.37** | 46.50 | 49.27 | 4.23 |
| ⊢ Deepseek-V3 | 74.80 | 90.12 | 93.14 | 91.86 | **82.79** | 96.31 | 91.34 | 98.95 | 34.15 | 50.08 | 15.77 |
| **Reasoning Models** | | | | | | | | | | | |
| ⊢ GPT-o1 | 55.12 | 92.37 | 93.48 | 90.78 | 81.40 | 96.67 | 91.44 | 99.32 | 56.91 | 39.67 | **3.41** |
| ⊢ Deepseek-R1 | 75.93 | 96.88 | 97.60 | 82.45 | 72.84 | 94.09 | 89.92 | 98.31 | 33.66 | 48.13 | 18.21 |
| ⊢ Gemini-2.5-Pro | 65.20 | **98.09** | **98.17** | 89.18 | 80.31 | 95.93 | 90.92 | 99.08 | 48.78 | 47.15 | 4.07 |
| ⊢ QwQ-32b | 65.53 | 91.33 | 93.94 | 88.49 | 77.49 | 95.88 | 90.89 | 99.01 | 45.69 | 46.99 | 7.32 |
| ⊢ Qwen3-235B-A22B | 73.17 | 93.07 | 94.74 | 86.26 | 76.80 | 94.87 | 90.31 | 98.62 | 36.42 | 51.38 | 12.20 |
| **Open-Source Models** | | | | | | | | | | | |
| **MOE Models** | | | | | | | | | | | |
| ⊢ Llama4 Maverick | 72.52 | 85.58 | 84.88 | 88.13 | 77.09 | 94.99 | 90.37 | 98.59 | 35.28 | 41.95 | 22.76 |
| ⊢ Llama4 Scout | 66.50 | 90.73 | 88.78 | 75.18 | 63.93 | 90.13 | 88.16 | 97.42 | 46.83 | 43.58 | 9.59 |
| ⊢ Qwen3-235B-A22B | 69.92 | 83.75 | 83.74 | 80.53 | 70.86 | 89.89 | 85.57 | 95.81 | 42.60 | 44.88 | 12.52 |
| ⊢ Qwen3-30B-A3B | 66.99 | 83.98 | 83.58 | 80.70 | 70.92 | 90.02 | 85.73 | 95.94 | 42.11 | 42.76 | 15.12 |
| **Dense Models** | | | | | | | | | | | |
| ⊢ Llama3-8B | 75.77 | 91.99 | 92.20 | 89.24 | 78.67 | 95.33 | 90.66 | 98.78 | 30.24 | 34.31 | 35.45 |
| ⊢ Llama3-3B | 75.28 | 82.61 | 82.11 | 90.17 | 80.66 | 95.59 | 90.57 | 98.70 | 30.89 | 33.98 | 35.12 |
| ⊢ Qwen3-8B | 57.40 | 71.17 | 72.52 | 91.71 | 81.06 | 96.39 | 90.92 | 99.15 | 52.52 | 38.21 | 9.27 |
| ⊢ Qwen3-4B | 58.54 | 62.01 | 65.20 | 89.05 | 78.03 | 94.66 | 89.90 | 98.82 | 48.78 | 35.45 | 15.77 |

Table 5: **Single-turn conversation performance metrics of various models across detoxification, fluency, content preservation, and sentiment polarity.** **Box** highlights the best performance for each metric among closed-source models, while Box highlights the best performance among open-source models.

| Model | Detox. Acc. | | | Fluency | | | | CntPres.↑ | Sentiment Polarity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S-CLS↑ | W-Clean↑ | S-Clean↑ | BLEU↑ | ChrF++↑ | BS_F1↑ | COM.↑ | | Toxic↓ | Neutral↑ | Polite ↓ |
| **Closed-Source Models** | | | | | | | | | | | |
| **Generation Models** | | | | | | | | | | | |
| ⊢ GPT-4o | 52.94 | 87.50 | 73.53 | 77.90 | 70.65 | 92.94 | 87.30 | 98.90 | 58.82 | 38.24 | 2.94 |
| ⊢ Qwen-Max | 55.88 | 86.25 | 70.59 | 88.61 | 80.67 | 95.94 | **88.86** | 99.10 | 64.71 | 32.35 | 2.94 |
| ⊢ Gemini-2.5-Flash | 47.06 | 86.25 | 76.47 | 90.16 | 79.82 | 95.78 | 88.23 | 99.06 | 64.71 | 32.35 | 2.94 |
| ⊢ Deepseek-V3 | 64.71 | 82.50 | 73.53 | **90.38** | **83.46** | **96.10** | 88.55 | 99.16 | 55.88 | 38.24 | 5.88 |
| **Reasoning Models** | | | | | | | | | | | |
| ⊢ GPT-o1 | 52.94 | 91.25 | 79.41 | 87.32 | 81.14 | 95.41 | 88.55 | **99.18** | 82.35 | 17.65 | **0.00** |
| ⊢ Deepseek-R1 | **76.47** | 92.50 | 85.29 | 78.60 | 73.32 | 91.85 | 85.93 | 98.21 | **29.41** | **55.88** | 14.71 |
| ⊢ Gemini-2.5-Pro | 55.88 | **98.75** | **97.06** | 80.67 | 74.39 | 93.02 | 87.49 | 98.74 | 64.71 | 32.35 | 2.94 |
| ⊢ QwQ-32b | 64.71 | 93.75 | 85.29 | 84.93 | 76.23 | 93.60 | 87.37 | 98.85 | 50.00 | 38.24 | 11.76 |
| ⊢ Qwen3-235B-A22B | 61.76 | 91.25 | 85.29 | 79.62 | 71.16 | 92.79 | 86.96 | 98.72 | 47.06 | 44.12 | 8.82 |
| **Open-Source Models** | | | | | | | | | | | |
| **MOE Models** | | | | | | | | | | | |
| ⊢ Llama4 Maverick | 41.18 | 72.50 | 58.82 | 77.85 | 67.26 | 92.55 | 87.45 | 98.55 | 67.65 | 26.47 | 5.88 |
| ⊢ Llama4 Scout | 50.00 | 85.00 | 70.59 | 76.89 | 67.48 | 93.03 | 86.87 | 99.11 | 73.53 | 20.59 | 5.88 |
| ⊢ Qwen3-235B-A22B | 61.76 | 81.25 | 70.59 | 82.37 | 75.16 | 94.27 | 87.08 | 99.01 | 61.76 | 29.41 | 8.82 |
| ⊢ Qwen3-30B-A3B | 41.18 | 67.50 | 52.94 | 90.56 | 83.14 | 96.02 | 88.18 | 99.32 | 67.65 | 32.35 | 0.00 |
| **Dense Models** | | | | | | | | | | | |
| ⊢ Llama3-8B | 61.76 | 82.50 | 70.59 | 58.59 | 60.22 | 79.14 | 70.28 | 84.29 | 52.94 | 8.82 | 38.24 |
| ⊢ Llama3-3B | 64.71 | 85.00 | 73.53 | 48.87 | 54.63 | 74.95 | 65.69 | 80.75 | 47.06 | 5.88 | 47.06 |
| ⊢ Qwen3-8B | 41.18 | 61.25 | 38.24 | 92.23 | 86.01 | 96.67 | 88.52 | 99.64 | 79.41 | 17.65 | 2.94 |
| ⊢ Qwen3-4B | 32.35 | 38.75 | 20.59 | 91.64 | 84.36 | 96.32 | 88.06 | 99.60 | 82.35 | 14.71 | 2.94 |

Table 6: **Multi-turn conversation performance metrics of various models across detoxification, fluency, content preservation, and sentiment polarity.** **Box** highlights the best performance for each metric among closed-source models, while Box highlights the best performance among open-source models.

| Model | Detox. Acc. | | | Fluency | | | | CntPres.↑ | Sentiment Polarity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S-CLS↑ | W-Clean↑ | S-Clean↑ | BLEU↑ | ChrF++↑ | BS_F1↑ | COM.↑ | | Toxic↓ | Neutral↑ | Polite ↓ |
| GPT-4o | 83.33 | 100.00 | 100.00 | 73.97 | 66.61 | 86.44 | 85.77 | 91.87 | 36.67 | 53.33 | 10.00 |
| Deepseek-R1 | 90.00 | 100.00 | 100.00 | 77.56 | 71.32 | 86.73 | 84.87 | 90.61 | 36.67 | 46.67 | 16.67 |
| Annonator1 | 96.67 | 100.00 | 100.00 | 83.75 | 73.24 | 90.99 | 87.98 | 92.61 | 3.33 | 90.00 | 6.67 |
| Annonator2 | 93.34 | 100.00 | 100.00 | 85.64 | 77.65 | 91.03 | 87.73 | 93.35 | 3.33 | 83.33 | 13.33 |

Table 7: **Performance between LLMs and human annotators** on detoxification, fluency, content preservation, and sentiment polarity.

| Setting | Detox. Acc. | | | Fluency | | | | CntPres.↑ | Sentiment Polarity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S-CLS↑ | W-Clean↑ | S-Clean↑ | BLEU↑ | ChrF++↑ | BS_F1↑ | COM.↑ | | Toxic↓ | Neutral↑ | Polite ↓ |
| **Qwen-Max** | | | | | | | | | | | |
| *single-sentence* | 93.04 | 97.81 | 97.68 | 48.14 | 37.66 | 86.06 | 87.68 | 91.44 | 12.58 | 77.29 | 10.13 |
| *+ single-turn conv.* | 85.98 | 96.15 | 95.96 | 76.52 | 69.41 | 90.22 | 89.19 | 94.64 | 22.87 | 66.18 | 10.95 |
| | (-7.06) | (-1.66) | (-1.72) | (+28.38) | (+31.75) | (+4.16) | (+1.51) | (+3.20) | (+10.29) | (-11.11) | (+0.82) |
| *+ single-turn conv. + homo* | 85.54 | 96.11 | 95.86 | 76.29 | 69.18 | 90.05 | 89.03 | 94.48 | 23.08 | 66.06 | 10.86 |
| | (-0.44) | (-0.04) | (-0.10) | (-0.23) | (-0.23) | (-0.17) | (-0.16) | (-0.16) | (+0.21) | (-0.12) | (-0.09) |
| *+ single-turn conv. + homo* | 84.69 | 95.98 | 95.66 | 76.07 | 68.95 | 89.89 | 88.89 | 94.34 | 24.05 | 65.18 | 10.78 |
| *+ emoji* | (-0.85) | (-0.13) | (-0.20) | (-0.22) | (-0.23) | (-0.16) | (-0.14) | (-0.14) | (+0.97) | (-0.88) | (-0.08) |
| *+ single-turn conv. + homo* | 84.06 | 95.62 | 95.12 | 76.82 | 69.82 | 90.02 | 88.89 | 94.45 | 24.94 | 64.46 | 10.60 |
| *+ emoji + multi-turn conv.* | (-0.63) | (-0.36) | (-0.54) | (+0.75) | (+0.87) | (+0.13) | (±0.00) | (+0.11) | (+0.89) | (-0.72) | (-0.18) |
| **GPT-o1** | | | | | | | | | | | |
| *single-sentence* | 90.48 | 96.28 | 95.97 | 42.88 | 31.88 | 84.41 | 87.12 | 92.20 | 19.66 | 74.11 | 6.23 |
| *+ single-turn conv.* | 75.31 | 95.01 | 94.07 | 76.97 | 69.01 | 89.67 | 88.97 | 95.25 | 35.63 | 59.34 | 5.02 |
| | (-5.17) | (-1.27) | (-1.90) | (+34.09) | (+37.13) | (+5.26) | (+1.85) | (+3.05) | (+15.97) | (-14.77) | (-1.21) |
| *+ single-turn conv. + homo* | 74.54 | 94.84 | 93.82 | 76.85 | 68.80 | 89.55 | 88.85 | 95.11 | 36.46 | 58.59 | 4.96 |
| | (-0.77) | (-0.17) | (-0.25) | (-0.12) | (-0.21) | (-0.12) | (-0.12) | (-0.14) | (+0.83) | (-0.75) | (-0.06) |
| *+ single-turn conv. + homo* | 73.32 | 94.61 | 93.50 | 76.76 | 68.59 | 89.47 | 88.73 | 95.02 | 37.52 | 57.62 | 4.86 |
| *+ emoji* | (-1.22) | (-0.23) | (-0.32) | (-0.09) | (-0.21) | (-0.08) | (-0.12) | (-0.09) | (+1.06) | (-0.97) | (-0.10) |
| *+ single-turn conv. + homo* | 72.88 | 94.48 | 93.19 | 77.41 | 69.53 | 89.60 | 88.73 | 95.11 | 38.50 | 56.75 | 4.76 |
| *+ emoji + multi-turn conv.* | (-0.44) | (-0.13) | (-0.31) | (+0.65) | (+0.94) | (+0.13) | (±0.00) | (+0.09) | (+0.98) | (-0.87) | (-0.10) |
| **Qwen3-235B-A22B** | | | | | | | | | | | |
| *single-sentence* | 89.26 | 97.62 | 97.56 | 48.42 | 35.95 | 85.76 | 86.81 | 91.42 | 19.54 | 69.11 | 11.36 |
| *+ single-turn conv.* | 80.96 | 95.06 | 95.26 | 77.67 | 67.92 | 89.86 | 88.46 | 94.57 | 29.43 | 58.72 | 11.85 |
| | (-8.30) | (-2.56) | (-2.30) | (+29.25) | (+31.97) | (+4.10) | (+1.65) | (+3.15) | (+9.89) | (-10.39) | (+0.49) |
| *+ single-turn conv. + homo* | 80.11 | 94.99 | 95.11 | 77.54 | 67.70 | 89.70 | 88.28 | 94.42 | 30.35 | 57.84 | 11.81 |
| | (-0.85) | (-0.07) | (-0.15) | (-0.13) | (-0.22) | (-0.16) | (-0.18) | (-0.15) | (+0.92) | (-0.88) | (-0.04) |
| *+ single-turn conv. + homo* | 78.65 | 94.86 | 94.88 | 77.43 | 67.47 | 89.60 | 88.15 | 94.32 | 31.67 | 56.83 | 11.50 |
| *+ emoji* | (-1.46) | (-0.13) | (-0.23) | (-0.11) | (-0.23) | (-0.10) | (-0.13) | (-0.10) | (+1.32) | (-1.01) | (-0.31) |
| *+ single-turn conv. + homo* | 78.28 | 94.34 | 94.34 | 77.73 | 68.08 | 89.70 | 88.12 | 94.43 | 32.33 | 56.23 | 11.44 |
| *+ emoji + multi-turn conv.* | (-0.37) | (-0.52) | (-0.54) | (+0.30) | (+0.61) | (+0.10) | (-0.03) | (+0.11) | (+0.66) | (-0.60) | (-0.06) |
| **Qwen3-8B** | | | | | | | | | | | |
| *single-sentence* | 89.87 | 94.85 | 94.26 | 41.84 | 29.90 | 83.02 | 85.41 | 90.38 | 15.51 | 71.06 | 13.43 |
| *+ single-turn conv.* | 75.94 | 84.09 | 84.94 | 77.96 | 68.24 | 88.75 | 87.77 | 94.14 | 31.38 | 56.97 | 11.65 |
| | (-13.93) | (-10.76) | (-9.32) | (+36.12) | (+38.34) | (+5.73) | (+2.36) | (+3.76) | (+15.87) | (-14.09) | (-1.78) |
| *+ single-turn conv. + homo* | 75.76 | 84.12 | 84.86 | 77.70 | 67.93 | 88.56 | 87.60 | 94.03 | 31.64 | 56.55 | 11.81 |
| | (-0.18) | (+0.03) | (-0.08) | (-0.26) | (-0.31) | (-0.19) | (-0.17) | (-0.11) | (+0.26) | (-0.42) | (+0.16) |
| *+ single-turn conv. + homo* | 75.16 | 84.32 | 84.95 | 77.35 | 67.55 | 88.34 | 87.44 | 93.87 | 32.06 | 56.24 | 11.70 |
| *+ emoji* | (-0.60) | (+0.20) | (+0.09) | (-0.35) | (-0.38) | (-0.22) | (-0.16) | (-0.16) | (+0.42) | (-0.31) | (-0.11) |
| *+ single-turn conv. + homo* | 74.42 | 83.45 | 83.93 | 82.04 | 70.07 | 89.96 | 87.87 | 94.00 | 33.10 | 55.40 | 11.50 |
| *+ emoji + multi-turn conv.* | (-0.74) | (-0.87) | (-1.02) | (+4.69) | (+2.52) | (+1.62) | (+0.43) | (+0.13) | (+1.04) | (-0.84) | (-0.20) |

Table 8: **Progressive breakdown of model performance with incrementally added task complexity.**

| Model | Detox. Acc. | | | Fluency | | | | CntPres.↑ | Sentiment Polarity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S-CLS↑ | W-Clean↑ | S-Clean↑ | BLEU↑ | ChrF++↑ | BS_F1↑ | COM.↑ | | Toxic↓ | Neutral↑ | Polite ↓ |
| Deepseek-R1 | **87.05** | 99.72 | 99.64 | 68.49 | 61.74 | 86.24 | 85.79 | 92.57 | **21.22** | 58.63 | 20.14 |
| Llama3-8B | 73.02 | 99.02 | 98.74 | 75.79 | 63.30 | 86.87 | 84.66 | **93.04** | 36.33 | 42.09 | 21.58 |
| Llama3-8B_1K | 83.81 | **100.00** | **100.00** | **76.83** | **66.50** | **87.39** | **85.81** | 92.77 | 21.94 | **63.85** | **14.21** |

Table 9: **Performance of fine-tuning LLaMA3-8B with 1K samples.** LLaMA3-8B_1K yields clear gains over the base model, notably +10.79 on S-CLS and +21.76 on neutral polarity, surpassing Deepseek-R1.