Reflect, Reason, Rephrase (R^3-Detox): An In-Context Learning Approach to Text Detoxification

Guillermo Villate-Castillo*
guillermo.villate@tecnalia.com
TECNALIA, Basque Research & Technology Alliance (BRTA)
Derio, Biscay, Spain

Javier Del Ser
javier.delser@ehu.eus
University of the Basque Country (UPV/EHU)
Leioa, Biscay, Spain

Borja Sanz Urquijo
borja.sanz@deusto.es
University of Deusto
Bilbao, Biscay, Spain

Abstract

Traditional content moderation, while effective in reducing toxicity through content removal or censoring, can discourage user participation by making them feel restricted or unfairly targeted, especially in nuanced discussions. Text detoxification offers a more constructive alternative by rephrasing offensive language into respectful forms. We propose R^3-Detox, a Reflect-Reason-Rephrase framework that structures detoxification into three steps within a single prompt. The model identifies potentially toxic elements guided by Shapley values to reduce fabricated predictions, evaluates overall toxicity, and then revises the text to eliminate toxicity while retaining meaning. We augment three offensive text paraphrasing datasets (ParaDetox, Parallel Detoxification, APPDIA) with explicit detoxification reasoning. Evaluated with in-context learning, R^3-Detox outperforms state-of-the-art methods, including instruction following models.

CCS Concepts

• **Computing methodologies → Natural language generation**.

Keywords

Text Detoxification, LLM, Self-Reflection, Reasoning

ACM Reference Format:
Guillermo Villate-Castillo, Javier Del Ser, and Borja Sanz Urquijo. 2025. Reflect, Reason, Rephrase (R^3-Detox): An In-Context Learning Approach to Text Detoxification. In *Proceedings of 12th ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT2025).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Disclaimer: Figures and examples shown in this manuscript may feature toxic language.

In the context of online social interactions, traditional methods such as content flagging and removal have been shown to mitigate harm to individuals or groups targeted by toxic speech, hate, and cyberbullying [5, 12, 27]. However, such approaches may suppress participation, limit diverse perspectives [21, 38], and disproportionately silence certain voices [42]. These issues highlight the need for approaches that mitigate toxicity while preserving open dialogue. Text detoxification addresses this by rephrasing offensive content into less harmful language without altering its meaning [28]. This promotes inclusive and constructive discussions; however, achieving fairness and accuracy requires sensitivity to context and societal norms [38].

Previous studies have explored the use of supervised generative models for paraphrasing offensive content, such as BART [28] and DialoGPT [1]. While performing well on certain metrics, including BLEU, BERTScore, and ROUGE, these models come with notable limitations. They require large amounts of labeled data, generalize poorly across domains, and may fail to fully eliminate toxic behavior [38]. To enhance adaptability, prior research has leveraged the In-Context Learning (ICL) [49] capabilities of Large Language Models (LLMs), showing promising results in both detoxification [19, 38] and synthetic data generation [31]. Additionally, recent approaches have leveraged the explanation capabilities of LLMs through Chain-of-Thought (CoT) prompting [44]. This method asks the model to explain why a sentence is toxic before performing detoxification, yielding interpretable and effective rewrites [23].

Expanding on recent advancements, we leverage LLMs to reconceptualize detoxification as a process of self-reflection [9] and abductive reasoning. To this end, we introduce the **R^3-Detox** framework, which emulates human cognitive processes to enhance detoxification quality [37]. This framework first guides the LLM in identifying potentially toxic words within a sentence using Shapley value-based explanations [30] extracted from toxicity detectors. Next, the framework instructs the LLM to analyze the underlying reasons for the sentence's toxicity based on these identified words. Finally, it directs the LLM to propose necessary modifications to neutralize the toxicity while preserving the original meaning, explaining how these changes promote a more inclusive and less toxic output.

To ensure the quality of reasoning, we resort to models trained in Self-Reflection (SR) and incorporate existing detoxification datasets, conditioning on human-generated rephrasing to maintain consistency across each step. R^3-Detox addresses several Research Questions (RQs) central to evaluating its effectiveness in text detoxification:

- RQ1: How do existing reasoning evaluation metrics, correlate with human evaluation in the task of text detoxification?
- RQ2: Can models trained in Self-Reflection reason in highly subjective tasks, such as text toxicity detection and detoxification?
- RQ3: Does R^3-Detox achieve better detoxification results than state-of-the-art techniques by using ICL with few-shot examples?

BDCAT2025, Nantes, France

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

Paper outline: Section 2 discusses related work, whereas Section 3 explains the methodology followed for the generation and validation of the proposed R^3-Detox framework. Section 4 outlines the evaluation metrics used and describes the human annotation process. Section 5 presents experimental results and summarizes key findings from the experiments. Finally, Section 7 summarizes the contributions and limitations of our study.

2 Related Work

Text style transfer modifies the style of a sentence while preserving its meaning, with text detoxification focusing specifically on rewriting toxic sentences into non-toxic ones [22]. Early approaches to detoxification include supervised and unsupervised methods. Supervised models such as COUNT [36] optimize for non-toxic outputs by penalizing toxicity, while unsupervised methods [25, 32] address the scarcity of reference detoxified data by using techniques like cycle consistency loss. However, both methods suffer from limited generalization, weak control over output quality, and poor interpretability. In response, some work has modeled detoxification as style-conditioned generation to improve control. For example, CondBERT [4] applies masked language modeling, whereas ParaGEDI [28] guides generation at the token level. These methods offer finer control but still struggle with consistency and transparency.

More recently, In-Context Learning (ICL) with LLMs has achieved better generalization [31, 38]. Beyond simple instructions, a rich array of prompting strategies has emerged to guide LLMs for detoxification. Som [38] notably demonstrate that LLMs prompted with carefully chosen demonstration pairs can even surpass fine-tuned models in toxicity reduction, highlighting the potential of pure ICL. Other advancements focus on self-correction and iterative refinement. For example, et al [15] introduce CRITIC, a framework where LLMs engage in verification-correction cycles by interacting with external toxicity scoring tools, using natural language feedback to revise their outputs. Similarly, Krishna [24] propose a two-stage multi-prompting approach where LLMs intrinsically reason step-by-step about toxicity and self-correct their generations based on their own assessment. Xu [46] introduce Perspective-Taking Prompting, which guides LLMs to reduce toxicity by invoking empathetic reasoning, prompting models to consider different perspectives for self-correction. Furthermore, GPT-DETOX by Pesaranghader [35] leverages effective zero-shot and few-shot prompting with novel example selection strategies to achieve high fluency and semantic similarity. More complex paradigms include the multi-expert prompting framework by Long [29], where multiple expert agents collaborate through dialogue to produce aggregated, safer responses. While prompt tuning, as explored by [20] with T5 models, shows promise, it exposes a critical trade-off between toxicity reduction and faithful content preservation.

To improve interpretability, some ICL approaches integrate CoT reasoning, prompting models to explain toxicity before rewriting [23, 47]. However, these explanations are often shallow and loosely connected to the detoxification itself, limiting their effectiveness. More recent work aims for deeper, more integrated explainability. For instance, Bhan [2] introduce CF-Detox$_{\text{tigtec}}$, which leverages local feature attribution (extracted via KernelSHAP) to pinpoint toxic words and counterfactual feature importance to assess the impact of edits, offering a more granular understanding of toxicity. Similarly, Lee [26] propose XDetox, which incorporates token-level explainability via DecompX for identifying toxic spans and guiding both the generation and selection of detoxified outputs. XDetox demonstrates how explanations can be intrinsically linked to the rewriting process. Similarly, Hallinan [16] developed MARCO, which implicitly identifies potentially toxic segments by computing Jensen-Shannon divergence between expert and anti-expert model predictions, providing an automated mechanism for localizing problematic content. Furthermore, to enhance the quality of CoT, the work in [23] introduces DetoxLLM, which utilizes LLMs to generate explicit, categorized explanations of toxicity for fine-tuning, yielding a structured and informative CoT reasoning.

Contribution. Building on these insights, we propose a structured reasoning framework that combines explanation-based reasoning and guided CoT within a single prompt. Unlike prior work, we show that SR models [9] are well suited for detoxification in an ICL setting. Our methodology identifies toxic elements, explains their harm, and reasons about targeted edits, improving transparency, control, and interpretability. By leveraging oracle-guided introspection during tuning, Self-Reflection models further enhance quality and reliability. This unified approach tightly integrates explanation, reasoning, and rewriting, advancing beyond existing methods.

3 Methodology

This section describes the methodology outlined in Figure 1, pausing at each of its components: the datasets in use (Subsection 3.1), the few-shot example generation pipeline (Subsection 3.2), and the ICL method used to validate the R^3-Detox framework (Subsection 3.3).

3.1 Datasets

To validate our R^3-Detox framework through ICL with few-shot examples, we leverage publicly available English text detoxification datasets. Specifically, we employ validated non-toxic paraphrases to ensure that the reasoning underlying the generation of detoxified text is grounded on the differences between the original and paraphrased sentences. The considered datasets include ParaDetox [28], APPDIA [1], and Parallel Detoxification [6]. These resources have three complementary purposes: (1) providing a benchmark for evaluation, (2) supplying few-shot demonstrations for ICL, and (3) supporting the assessment and selection of models most suitable for the reasoning task. By leveraging datasets annotated by human experts, we ensure high-quality guidance for the model's reasoning process, enabling the framework to learn how to achieve effective text detoxification.

The reflection component of the R^3-Detox framework evaluates whether potentially toxic words convey harmful meaning within their local context. To validate this capability, we incorporate non-toxic data from the Jigsaw Unintended Bias dataset [3], selecting only comments annotated by at least 10 raters to ensure reliability and exclude any residual toxicity. In total, we collected 14,969 toxic and 14,969 non-toxic sentences, creating a class-balanced dataset which is hereafter coined as the Reflect, Reason, Paraphrase (RRP) dataset. This dataset is used for few-shot demonstration generation and for evaluating model suitability in the reasoning step.

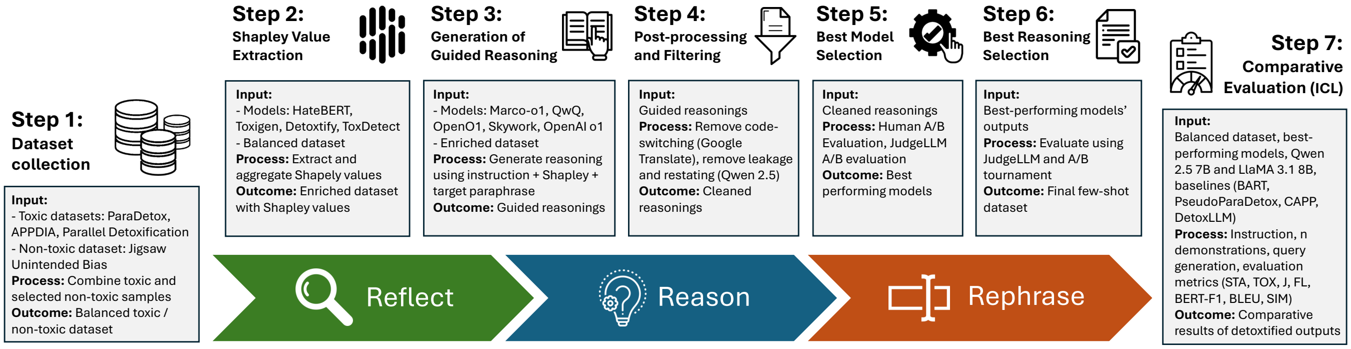

Step 1: Dataset collection
Input: - Toxic datasets: ParaDetox, APPDIA, Parallel Detoxification - Non-toxic dataset: Jigsaw Unintended Bias
Process: Combine toxic and selected non-toxic samples
Outcome: Balanced toxic / non-toxic dataset
Step 2: Shapley Value Extraction
Input: - Models: HateBERT, Toxigen, Detoxtify, ToxDetect - Balanced dataset
Process: Extract and aggregate Shapely values
Outcome: Enriched dataset with Shapley values
Step 3: Generation of Guided Reasoning
Input: - Models: Marco-o1, QwQ, OpenO1, Skywork, OpenAI o1 - Enriched dataset
Process: Generate reasoning using instruction + Shapley + target paraphrase
Outcome: Guided reasonings
Step 4: Post-processing and Filtering
Input: Guided reasonings
Process: Remove code-switching (Google Translate), remove leakage and restating (Qwen 2.5)
Outcome: Cleaned reasonings
Step 5: Best Model Selection
Input: Cleaned reasonings
Process: Human A/B Evaluation, JudgeLLM A/B evaluation
Outcome: Best performing models
Step 6: Best Reasoning Selection
Input: Best-performing models' outputs
Process: Evaluate using JudgeLLM and A/B tournament
Outcome: Final few-shot dataset
Step 7: Comparative Evaluation (ICL)
Input: Balanced dataset, best-performing models, Qwen 2.5 7B and LlaMA 3.1 8B, baselines (BART, PseudoParaDetox, CAPP, DetoxLLM)
Process: Instruction, n demonstrations, query generation, evaluation metrics (STA, TOX, J, FL, BERT-F1, BLEU, SIM)
Outcome: Comparative results of detoxtified outputs
Reflect
Reason
Rephrase

Figure 1: Methodology implemented by the R^3-Detox framework. We first preprocess the datasets (Section 3.1) by extracting Shapley values from toxicity detectors. Guided reasoning is then generated using SR models, ensuring no code-switching or data leakage so that the final non-toxic paraphrase is not explicitly present before detoxification. We evaluate models, select the best reasoning for each comment, and validate few-shot examples by comparing them to avant-garde detoxification techniques using ICL.

While our experiments focus on single-sentence inputs due to limited resources and the scarcity of large-scale contextual detoxification datasets, we acknowledge that broader discourse context plays an important role in shaping the reasoning process for generating accurate non-toxic paraphrases. This issue is further examined in Section 6.3.2, where we discuss the influence of context on reasoning quality within detoxification tasks.

3.2 Generation of Few-Shot Examples

To guide the model toward correct reasoning and detoxification behavior, we provide few-shot examples illustrating desired outcomes, following Som [38], who show that few-shot prompting improves text detoxification in LLMs. Leveraging these examples also helps the model better follow the self-reflection and reasoning process. Building on this, we employ Shapley values from toxicity detectors to constrain possible fabricated predictions during reasoning by incorporating prior knowledge about potentially toxic words. Despite their utility, these toxicity models are not without limitations, including biases [48], challenges in detecting implicit toxicity [18], and generalization issues [17]. To mitigate these limitations, we aggregate Shapley values by selecting tokens flagged as potentially toxic by multiple models in agreement. Specifically, we use Toxigen HateBERT and Toxigen RoBERTa [18] for implicit toxicity detection; Toxic BERT and Unbiased Toxic RoBERTa [17] to address generalization; and ToxDetect RoBERTa Large [48] to reduce bias. Further details on this process are provided in Appendix A.

For the generation of abductive reasoning, we use several open-source models: Marco-o1 [14], QwQ Preview [41], OpenO1 LLaMA 8B v0.1 [34], and Skywork-o1-Open-Llama-3.1-8B [33], as well as the private OpenAI o1 [10]. The reasoning generation is guided by the Shapley values and constraint by the non-toxic paraphrases in the dataset by providing the possible toxic words and the final non-toxic paraphrase in the prompt as context.

While the models generate high-quality reasoning, we encountered several issues along the way, including code-switching, a tendency to restate the provided non-toxic paraphrase, and instances where the OpenAI moderation tool flagged our queries as toxic. To overcome these problems, we employ the Qwen 2.5 32B model [40] to identify and eliminate unwanted behaviors, resort to the Google Translate API to mitigate language-mixing problems, and apply the latest jail-breaking technique introduced in et al [7].

Finally, to construct the dataset, we select the best reasoning outputs based on the JudgeLLM [8] evaluation model, which is later described in Section 4. JudgeLLM has the highest correlation with human annotations, as is later empirically shown in Section 6.1. We further refine our selection by leveraging the best-performing model identified from our experiments corresponding to RQ2 (Section 6.2). The final reasoning outputs are determined by an “A vs. B” tournament evaluation using the JudgeLLM model (Section 4.1).

3.3 In-Context Learning

ICL is an approach that consists of three components: 1) an instruction I explaining the task to be performed, 2) a set of n demonstrations from the Reflect, Reason, Paraphrase generated dataset, and 3) a query, which is the toxic sentence that needs to be rewritten. In our framework, we adopt the methodology recently proposed in Som [38], which selects the most similar sentences based on a content similarity model, all-mpnet-base-v2 [39]. Prompts used in our framework are presented in Appendix A.

4 Evaluation Framework

In this section, we present the evaluation framework used to assess both the generated reasoning and the resulting non-toxic paraphrases. For each task, we describe the automatic metrics employed and the human evaluation procedures. All human evaluations were conducted by three volunteers (two female, one male), aged between 25 and 31, from Western Europe. Additional implementation details and examples are provided in Appendix B.

4.1 Reasoning Evaluation

To evaluate the quality of the generated reasoning, we employ the ROSCOE metric suite [11], which measures different aspects of reasoning quality through multiple sub-metrics: ROSCOE-SA (semantic alignment), ROSCOE-SS (semantic similarity), ROSCOE-LC (grammatical acceptability), Discourse Representation (contradiction probability per reasoning step), and Coherence (maximum contradiction probability between reasoning steps). In addition, we use JudgeLLM [8], a model built upon Vicuna and trained on large-scale datasets of LLM-generated responses annotated by GPT-4. JudgeLLM provides an automated assessment of response quality and has demonstrated over 90% agreement with human judgments in certain tasks, making it a suitable complementary evaluator.

To analyze the alignment between automatic metrics and human preferences, we adopt an *"A vs. B"* comparative evaluation framework to rank models according to their reasoning quality in the R^3-Detox task. For each non-toxic paraphrase, reasoning outputs from n models are compared pairwise in a round-robin tournament. Each comparison awards 1 point to the preferred model and 0 points to the other; ties receive no points. The total number of pairwise comparisons is given by:

$$\binom{n}{2} \cdot h = \frac{n!}{2!(n-2)!} \cdot h, \tag{1}$$

where h denotes the number of evaluated paraphrases. Due to the effort required for human evaluation, we assess a total of 20 instances (5 per dataset), resulting in $\binom{5}{2} \cdot 20 = 200$ tournaments. Each tournament is independently judged by three annotators, yielding an inter-annotator agreement of 0.183 (Fleiss' Kappa). When no clear preference emerges (i.e., results include both "win A" and "win B" votes or ties), the pair is classified as a tie.

4.2 Paraphrase Evaluation

To assess the generated non-toxic paraphrases, we use a combination of standard automatic metrics drawn from previous work [23, 28]: Style Transfer Accuracy (STA) [28], BERTScore [13], Content Preservation (SIM) [45], Fluency (FL) [43], Joint Score (J) [28], and Toxicity (Tox) [38]. These metrics collectively measure how well the model achieves detoxification, preserves content, and maintains fluency.

Similarly to the reasoning evaluation, we also use JudgeLLM to assess the overall quality of the generated paraphrases. To analyze the alignment between human judgments and automatic metrics, we propose a *triplet elimination tournament* for ranking model outputs. In this setting, models are randomly grouped into triplets for each toxic input, and the best output from each group advances to the next round. The number of comparisons is given by $m \cdot h = \frac{n-1}{2} \cdot h$, where m denotes the number of triplet evaluations per comment and h the number of evaluated samples. This approach reduces annotation overhead while allowing finer-grained comparisons among paraphrases. Final rankings are aggregated using the Borda count method across all data points.

For our experiments, we evaluate 51 toxic comments (17 each from APPDIA, ParaDetox, and Parallel Detoxification). The evaluation proceeds in two phases:

- *Phase 1:* All three annotators independently evaluate 153 common triplets.
- *Phase 2:* Based on the first phase's results, each annotator evaluates an additional 51 triplets, where the top-performing models from Phase 1 compete.

The overall inter-annotator agreement for Phase 1 is Fleiss' Kappa = 0.09, reflecting the inherent subjectivity of paraphrase quality evaluation, where subtle differences in fluency, tone, or semantic preservation can lead to diverse interpretations.

5 Experimental Setup

We now present the experimental setup for the generation of the few-shot examples (Subsection 5.1) and explain the ICL methodology performed to evaluate the R^3-Detox framework (Subsection 5.2).

5.1 Few-Shot Synthetic Data Generation

To validate the capabilities of the Self-Reflection models presented in Section 3.2 when used in R^3-Detox, we adopt the "A vs. B" evaluation method outlined in Section 4.1, applying it to 20 comments manually analyzed by three annotators. Additionally, we perform the same analysis using the following metrics: ROSCOE-SA, ROSCOE-SS, ROSCOE-LC, discourse representation, coherence, and the JudgeLLM model, all within the context of the same tournament ranking. Finally, we compute the Spearman's rank correlation between the metrics, LLM evaluations, and human annotations.

5.2 In-Context Learning

We perform ICL using five open-source models: Marco-o1, OpenO1, and QwQ Preview, selected based on their strong reasoning generation performance, as evaluated in Section 6.2. Additionally, we include the Llama 3.1 8B and Qwen 2 7B instruct models to compare them against OpenO1 and Marco-o1, respectively. Marco-o1 and OpenO1 were fine-tuned on the self-reflection task derived from these instruction models.

For our experiments, we select the following numbers of examples: [0, 1, 2, 3, 5, 7, 10]. This selection is constrained by the maximum context length of 8,096 tokens imposed by the limited computational resources available for the study. To validate our approach, we compare our results against human-annotated non-toxic paraphrases from each dataset, as well as several baseline methods: DetoxLLM from Khondaker [23], the BART model trained for the detoxification task by ParaDetox [28], and the ICL method introduced in PseudoParaDetox [31] for synthetic data generation, which utilized the dolphin-2.9-llama3-8b ablated model. We also consider the first ICL method proposed recently for detoxification in CAPP [38].

We use the metrics mentioned in Section 4.2 to compare the performance of each ICL model on the same dataset. However, for the last approach (CAPP), a direct comparison is not possible, as GPT-3.5 models are no longer available and no code is provided. Instead, we use the example-generated dataset available in their repository[1].

6 Results and Discussion

This section presents the results for each of the research questions posed in Section 1. Code and dataset are available at Github[2].

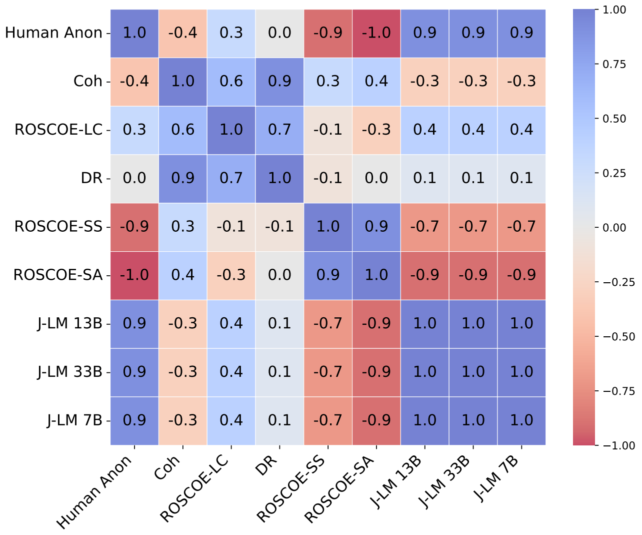


Figure 2: Spearman's rank correlation coefficients among metrics. J-LM means JudgeLM, DR is Discourse Representation, and Coh denotes Coherence.

6.1 RQ1: Correlation of Reasoning Evaluation Metrics and Human Annotated Rankings

Figure 2 illustrates Spearman's rank correlation (ρ) among all metrics, including the majority vote of the human annotations introduced in Section 4.1. All rankings are computed using the "A vs B" pairwise comparison described in Section 4.1, calculated across the 20 comments in an overall 200 pair tournament. The correlation matrix depicts the correlation of the ranks assigned to the aggregated ranking of the pairwise tournaments.

The results in this matrix show a strong correlation between JudgeLLM models and the human annotation majority vote, with the three variants achieving a ρ of 0.90. Although each model shows a similar Spearman rank correlation, when examining the aggregated correlation at the instance level we observe aggregated Spearman rankings using Fisher's method of 0.75, 0.736, and 0.769, with p-values of 0.08, 0.02, and 0.01 for JudgeLLM 7B, 13B, and 33B, respectively. These aggregated values evince that only the JudgeLLM 33B and 13B models demonstrate a statistically significant difference from the null hypothesis, as indicated by their p-values of 0.01 and 0.02. On the other hand, the JudgeLLM 7B model, with a p-value of 0.08, does not show significant deviation from randomness, highlighting the variability in performance based on model size and complexity. Overall, the JudgeLLM 33B shows the best correlation with human annotators.

In contrast, the ROSCOE metrics correlate very poorly, with the best among the metrics being ROSCOE-LC, which has a ρ of 0.3. These results are expected, as these metrics only account for semantic consistency, logicality, informativeness, fluency, and factuality of the generated reasoning, rather than its content itself.

Rank	Human	*Score*	JudgeLM 33B	*Score*
1	QwQ Preview	43	QwQ Preview	57
2	OpenO1	42	Marco-o1	50
3	Marco-o1	39	OpenO1	49
4	OpenAI o1	27	OpenAI o1	25
5	Skywork-o1	2	Skywork-o1	5

Table 1: Human and JudgeLLM rankings, including the final scores obtained from the pairwise comparisons.

To end with the correlation analysis in response to RQ1, Table 1 presents a comparison of the final rankings based on the scoring methodology explained in Section 4.1 between the performed manual and automatic pairwise comparisons. As we observe in this table, the overall ranking is practically the same across models. A exception are the rankings of Marco-o1 and OpenO1—models which, in both cases, score practically identical values relative to each other. By examining the obtained scores, we can observe that JudgeLLM tends to be more extreme, generating ties for 24 out of 200 pairwise comparisons, whereas the human evaluation is more lenient, generating ties in 47 out of 200 comparisons. Although we observe a disparity in the scores in the top part of the ranking, the lower ranks are similar in both cases. Regarding Skywork-o1, the poor results reflect the tendency of this model to generate code instead of resolving the task with a plain text output. The OpenAI-o1 model perform worse because they only provide the final result, lacking intermediate reasoning. Additionally, bypassing the OpenAI Moderation tool and using the prompt injection technique [7] introduce noise, degrading the quality of the produced output.

6.2 RQ2: Can Self-Reflection models Reflect, Reason and Rephrase?

During the annotation process, we evaluate the correlation of the metrics, including those based on LLM, as shown in Figure 2. Most Self-Reflection models successfully identify toxic words in a sentence, assess overall toxicity, and suggest necessary changes. For non-toxic sentences, they explain why no toxicity is present, analyzing why potential toxic words are not harmful. Given that Marco-o1, OpenO1, and QwQ Preview perform best in the human annotation process, we further differentiate their capabilities by generating reasoning outputs for the entire dataset. To validate results, we apply the same "A vs. B" tournament from Section 4.1, using JudgeLLM 33B due to its high correlation with human evaluations.

JudgeLLM performs 89,856 pairwise evaluations on 29,952 dataset samples. OpenO1 ranks highest, winning 32,518 out of 86,526 possible scores, followed by QwQ Preview with 27,104 and Marco-o1 with 23,687. However, since the top model wins only 39.03% of the total pairs, all three models demonstrate strong performance in the Reflect, Reason, and Paraphrase reasoning tasks. The only limitation is the code-switching behavior observed in Marco-o1 and QwQ Preview, which we address by automatically translating the text from Chinese to English. Even when the text contains mixed languages, it remains coherently generated by the models.

[1] CAPP GitHub: https://github.com/anirudhsom/CAPP-Dataset [accessed: 02/13/2025].
[2] R^3-Detox: https://github.com/TheMrguiller/R-3-Detox [accessed: 10/22/2025]

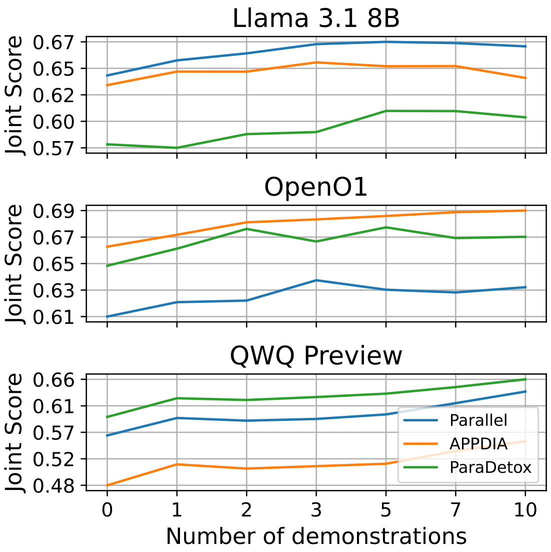


Figure 3: Joint Score versus the number of examples.

6.3 RQ3: Reflect, Reason, Rephrase ICL versus State-of-the-Art

In this section we first analyze the impact of the number of demonstrations on the results (Section 6.3.1). Section 6.3.2 compares R^3-detox with state-of-the-art techniques. Finally, Section 6.3.3 examines the correlation between the metrics introduced in Section 4.2.

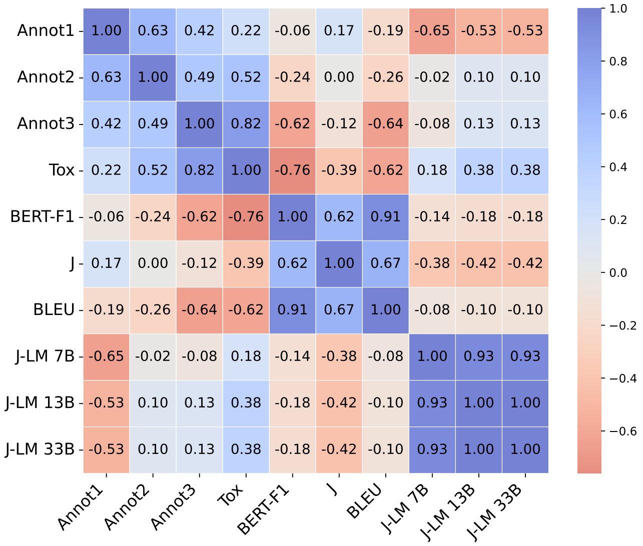
6.3.1 Importance of Number of Demonstrations. Figure 3 shows the impact of introducing demonstrations on the J metric across a set of models to highlight the variability of this impact. We observe that after two examples are provided, the J generally improves from the zero setting, following the Reflect, Reason, Rephrase reasoning methodology. Furthermore, while increasing the number of examples improves performance, the effect varies depending on the model. For example, the QwQ Preview model tends to improve as more demonstrations are provided, contrarily to Llama 3.1 8B and OpenO1. As more sentences are provided, the output of these latter models become less similar to the query, potentially inducing noise on the model's performance.

Notably, models fine-tuned in Self-Reflection, such as OpenO1 derived from the base Llama 3.1 8B model, appear to be more capable, adjusting better to the task of detoxification compared to the base instruction-tuned models.

6.3.2 Comparison with state-of-the-art approaches. The comparison between R^3-Detox and state-of-the-art models is shown in Table 2. The goal of a detoxification model is to generate paraphrases that preserve meaning and fluency while effectively reducing toxicity. Similarity metrics such as BERT-F1, BLEU, and SIM assess content preservation, but can be misleading for detoxification. For example, the BART and PseudoParaDetox baselines achieve the highest similarity scores, yet also exhibit higher toxicity and lower STA than the gold standard, suggesting that they often retain toxic content to boost similarity at the expense of detoxification. This highlights that detoxification-specific metrics, STA, Tox, and J are more meaningful indicators. Our R^3-Detox framework and DetoxLLM achieve the lowest toxicity and highest STA, demonstrating the value of incorporating explicit reasoning, while the superior J scores further confirm our method's strong overall balance between detoxification and content preservation.

All models except BART also outperform the gold standard in terms of FL, likely because they correct minor grammatical errors in the original sentences. Notably, QwQ Preview achieves the lowest toxicity at the cost of a significantly lower similarity, likely due to over-contextualizing or over-correcting the input (as discussed in Section 7). Overall, R^3-Detox achieves the best trade-off between meaning preservation and toxicity reduction, highlighting the value of structured reasoning for effective and controlled detoxification.

6.3.3 Correlation Between Metric Rankings and Annotator Evaluations. Figure 4 shows the correlation between the metric rankings (Section 4.2) and those of the three annotators. Inter-annotator agreement is low, with Spearman's correlations below 0.2 and p-values of 0.09, 0.27, and 0.18 for Annotator pairs (1–2, 1–3, and 2–3), indicating inconsistency in their rankings.



	Annot1	Annot2	Annot3	Tox	BERT-F1	J	BLEU	J-LM 7B	J-LM 13B	J-LM 33B
Annot1	1.00	0.63	0.42	0.22	-0.06	0.17	-0.19	-0.65	-0.53	-0.53
Annot2	0.63	1.00	0.49	0.52	-0.24	0.00	-0.26	-0.02	0.10	0.10
Annot3	0.42	0.49	1.00	0.82	-0.62	-0.12	-0.64	-0.08	0.13	0.13
Tox	0.22	0.52	0.82	1.00	-0.76	-0.39	-0.62	0.18	0.38	0.38
BERT-F1	-0.06	-0.24	-0.62	-0.76	1.00	0.62	0.91	-0.14	-0.18	-0.18
J	0.17	0.00	-0.12	-0.39	0.62	1.00	0.67	-0.38	-0.42	-0.42
BLEU	-0.19	-0.26	-0.64	-0.62	0.91	0.67	1.00	-0.08	-0.10	-0.10
J-LM 7B	-0.65	-0.02	-0.08	0.18	-0.14	-0.38	-0.08	1.00	0.93	0.93
J-LM 13B	-0.53	0.10	0.13	0.38	-0.18	-0.42	-0.10	0.93	1.00	1.00
J-LM 33B	-0.53	0.10	0.13	0.38	-0.18	-0.42	-0.10	0.93	1.00	1.00

Figure 4: Spearman's rank correlation coefficients among metrics used to evaluate detoxification results.

Rather than interpreting this low agreement as a flaw in the annotation design, we regard it as a meaningful outcome that underscores the ambiguity and variability of human judgment in this nuanced language generation task, especially in cases where baseline models already perform well. As further detailed in Section 7, the subjective nature of toxicity perception, combined with subtle semantic and pragmatic differences in paraphrases, makes consistent agreement challenging to achieve.

7 Conclusions

In this paper, we have introduced a new framework, coined as *Reflect, Reason, and Rephrase* (R^3-Detox), which transforms the text

Dataset	Method	BERT-F1 ↑	BLEU ↑	SIM ↑	FL ↑	STA ↑	J ↑	Tox ↓
APPDIA	Original Sentence	-	-	-	-	-	-	0.748
	Gold-Standard	0.954	0.516	0.784	0.912	0.887	0.634	0.134
	BART	0.972	0.668	0.881	0.861	0.808	0.612	0.221
	DetoxLLM	0.925	0.214	0.654	0.967	0.922	0.583	0.059
	PseudoParaDetox	0.95	0.442	0.772	0.949	0.778	0.57	0.203
	CAPP*	0.955	0.521	0.808	0.971	0.898	0.704	0.117
	(R^3-Detox) Marco-o1	0.936 ± 0.002	0.336 ± 0.013	0.692 ± 0.012	0.948 ± 0.007	0.925 ± 0.009	0.607 ± 0.01	0.077 ± 0.01
	(R^3-Detox) Qwen 2.5 7B	0.926 ± 0.011	0.284 ± 0.068	0.649 ± 0.065	0.932 ± 0.026	0.958 ± 0.012	0.577 ± 0.04	0.048 ± 0.015
	(R^3-Detox) OpenO1	0.934 ± 0.003	0.324 ± 0.015	0.686 ± 0.016	0.963 ± 0.006	0.948 ± 0.008	0.627 ± 0.008	0.055 ± 0.007
	(R^3-Detox) Llama 3.1 8B	0.93 ± 0.005	0.326 ± 0.035	0.653 ± 0.031	0.945 ± 0.015	0.959 ± 0.01	0.593 ± 0.014	0.053 ± 0.01
	(R^3-Detox) QwQ Preview	0.909 ± 0.004	0.183 ± 0.022	0.529 ± 0.027	0.987 ± 0.002	0.986 ± 0.004	0.515 ± 0.024	0.02 ± 0.004
ParaDetox	Original Sentence	-	-	-	-	-	-	0.892
	Gold-Standard	0.951	0.47	0.813	0.805	0.943	0.617	0.0763
	BART	0.961	0.555	0.862	0.831	0.924	0.662	0.091
	DetoxLLM	0.922	0.203	0.68	0.967	0.951	0.625	0.033
	PseudoParaDetox	0.943	0.394	0.799	0.923	0.859	0.633	0.117
	CAPP*	0.955	0.486	0.849	0.939	0.945	0.754	0.06
	(R^3-Detox) Marco-o1	0.94 ± 0.002	0.366 ± 0.012	0.771 ± 0.008	0.904 ± 0.002	0.936 ± 0.006	0.652 ± 0.009	0.064 ± 0.003
	(R^3-Detox) Qwen 2.5 7B	0.931 ± 0.011	0.32 ± 0.072	0.734 ± 0.066	0.903 ± 0.033	0.969 ± 0.009	0.641 ± 0.033	0.036 ± 0.012
	(R^3-Detox) OpenO1	0.938 ± 0.002	0.349 ± 0.016	0.767 ± 0.014	0.931 ± 0.007	0.947 ± 0.003	0.677 ± 0.009	0.054 ± 0.001
	(R^3-Detox) Llama 3.1 8B	0.936 ± 0.005	0.358 ± 0.037	0.747 ± 0.03	0.919 ± 0.02	0.967 ± 0.003	0.663 ± 0.011	0.043 ± 0.005
	(R^3-Detox) QwQ Preview	0.921 ± 0.005	0.247 ± 0.026	0.663 ± 0.029	0.972 ± 0.006	0.979 ± 0.006	0.631 ± 0.02	0.025 ± 0.004
Parallel	Original Sentence	-	-	-	-	-	-	0.836
	Gold-Standard	0.934	0.369	0.724	0.801	0.92	0.533	0.09
	BART	0.966	0.63	0.875	0.876	0.794	0.609	0.165
	DetoxLLM	0.922	0.203	0.68	0.967	0.951	0.625	0.033
	PseudoParaDetox	0.946	0.423	0.807	0.929	0.781	0.585	0.163
	(R^3-Detox) Marco-o1	0.937 ± 0.002	0.368 ± 0.014	0.756 ± 0.011	0.927 ± 0.005	0.917 ± 0.005	0.643 ± 0.008	0.074 ± 0.009
	(R^3-Detox) Qwen 2.5 7B	0.928 ± 0.011	0.315 ± 0.073	0.711 ± 0.067	0.921 ± 0.028	0.955 ± 0.014	0.624 ± 0.036	0.044 ± 0.014
	(R^3-Detox) OpenO1	0.934 ± 0.002	0.343 ± 0.017	0.746 ± 0.015	0.953 ± 0.01	0.935 ± 0.006	0.665 ± 0.009	0.061 ± 0.004
	(R^3-Detox) Llama 3.1 8B	0.933 ± 0.005	0.359 ± 0.038	0.725 ± 0.026	0.937 ± 0.021	0.95 ± 0.011	0.646 ± 0.007	0.056 ± 0.007
	(R^3-Detox) QwQ Preview	0.916 ± 0.005	0.236 ± 0.026	0.627 ± 0.031	0.981 ± 0.003	0.974 ± 0.008	0.599 ± 0.024	0.028 ± 0.007

Table 2: Quantitative assessment of different LLMs based on our R^3-Detox approach and comparison against state-of-the-art detoxification techniques. The toxicity of the original sentence is provided, and the dataset's non-toxic paraphrase metric is used as the Gold Standard. Mean and standard deviation (std) values are computed over the different few-shot values. Best and worst results are shaded in blue and red, respectively. CAPP* values are based on the small subset available in their GitHub repository [38].

detoxification task into a three-step reasoning process. Through our methodology, we generate a first dataset that explains the intermediate analysis required to produce a non-toxic final sentence using open-source Self-Reflection models. To validate the quality of the generated reasoning, we then use human evaluation to assess how well human annotations correlate with the ROSCOE metric and JudgeLLM evaluations. Based on this correlation, we select the most suitable metric for evaluating the intermediate analysis generated by the Self-Reflection models, eventually determining the best-performing approach. This dataset is later used as a few-shot example set in ICL to validate our R^3-Detox framework. It helps generate non-toxic paraphrases from toxic inputs by explaining why a given sentence is toxic and what changes are needed to make it non-toxic

In the experiments discussed in this study, we have observed that the JudgeLLM evaluation model exhibits a high correlation with human annotation, making it a suitable tool for assessing the generated reasoning. We have also demonstrated that R^3-Detox, based on the generated demonstrations, outperforms existing state-of-the-art techniques, producing rephrased sentences that are less toxic and retain their original meaning. Finally, we have assessed and concluded that the manual annotation of the generated paraphrases is a complex, subjective task that requires standardized guidelines to improve the consistency of annotations.

Limitations

While our R^3-Detox framework makes meaningful contributions to detoxification, several limitations remain. First, the R^3-Detox framework inherits weaknesses from its components. Explanations from toxicity detectors can still reflect bias (e.g., flagging *"gay"* as toxic), and thus Shapley values may not always be reliable. Similarly, Self-Reflection models may over-contextualize input, sometimes altering meaning–as when *"A sociopathic idiot is trying to reassure us"* becomes *"Given their history of unreliable behavior, their attempts to reassure us are met with skepticism"* by QwQ Preview model. Second, our evaluation dataset is limited. It should be expanded to better capture implicit toxicity (e.g., sarcasm, coded language), conversational contexts, and non-detoxifiable extreme cases absent from current benchmarks. Broader evaluation would clarify how the framework performs in such settings. Third, detoxification techniques can still produce subtle toxicity that escapes classifiers, raising the risk of malicious use. Finally, annotation introduces subjectivity. Perceptions of toxicity vary with cultural and personal context, and our annotator pool was relatively small and regionally homogeneous, which may limit generalizability. Low inter-annotator agreement highlights the difficulty of evaluating toxicity rather than flaws in annotation design. Specific challenges include synonym variability, where sentences differing only by synonyms (e.g., “not very bright” vs. “not very smart”) can yield inconsistent annotations depending on perceived offensiveness; ambiguity in meaning preservation, where keeping the original intent in the paraphrased text is difficult and even minimal edits can leave subtle connotations (e.g., “I hope the bastard suffered” vs. “I hope the person suffered”); and slang and niche terms, which are hard to interpret consistently and introduce subjectivity in toxicity evaluation. Overall, these limitations reflect the persistent complexity of paraphrasing toxic content. Future work should broaden datasets, diversify annotators, and refine evaluation protocols to better capture the nuanced and subjective nature of toxicity.

Acknowledgments

Guillermo Villate-Castillo acknowledges the funding support provided by the European Union for research in the defense sector under the Future Combat Air System (FCAS) project, Phase 1B (RFP_ID: CYB_03), within the Cyber Core Services program.

References

[1] Katherine et al Atwell. 2022. APPDIA: A Discourse-aware Transformer-based Style Transfer Model for Offensive Social Media Conversations. In Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea, 6063–6074.

[2] Milan et al Bhan. 2024. Mitigating Text Toxicity with Counterfactual Generation. arXiv preprint arXiv:2405.09948 (2024).

[3] cjadams et al. 2019. Jigsaw Unintended Bias in Toxicity Classification.

[4] David et al Dale. 2021. Text Detoxification using Large Pre-trained Neural Models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic, 7979–7996.

[5] Thomas et al Davidson. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In Proceedings of the 11th International AAAI Conference on Web and Social Media (Montreal, Canada) (ICWSM ’17). 512–515.

[6] Daryna Dementieva et al. 2021. Crowdsourcing of Parallel Corpora: the Case of Style Transfer for Detoxification. In CSW@VLDB.

[7] John Hughes et al. 2024. Best-of-N Jailbreaking. arXiv:2412.03556 [cs.CL]

[8] Lianghui Zhu et al. 2025. JudgeLM: Fine-tuned Large Language Models are Scalable Judges. In The Thirteenth International Conference on Learning Representations.

[9] Ming Li et al. 2023. Reflection-Tuning: Recycling Data for Better Instruction-Tuning. In NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following.

[10] OpenAI et al. 2024. OpenAI o1 System Card. arXiv:2412.16720 [cs.AI]

[11] Olga Golovneva et al. 2022. ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning. CoRR abs/2212.07919 (2022).

[12] Robert Gorwa et al. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society 7, 1 (2020), 2053951719897945.

[13] Tianyu Gao et al. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. CoRR abs/2104.08821 (2021). arXiv:2104.08821

[14] Yu Zhao et al. 2024. Marco-o1: Towards Open Reasoning Models for Open-Ended Solutions. arXiv:2411.14405 [cs.CL]

[15] Zhibin Gou et al. 2024. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In The Twelfth International Conference on Learning Representations.

[16] Skyler et al Hallinan. 2023. Detoxifying Text with MaRCo: Controllable Revision with Experts and Anti-Experts. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Toronto, Canada, 228–242.

[17] Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

[18] Thomas et al Hartvigsen. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, 3309–3326.

[19] Xinlei et al He. 2024. You Only Prompt Once: On the Capabilities of Prompt Learning on Large Language Models to Tackle Toxic Content . In 2024 IEEE Symposium on Security and Privacy (SP). IEEE Computer Society, Los Alamitos, CA, USA, 770–787.

[20] Xinlei et al He. 2024. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In 2024 IEEE Symposium on Security and Privacy (SP). IEEE, 770–787.

[21] Shagun et al Jhaver. 2019. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 192 (Nov. 2019), 33 pages.

[22] Di et al Jin. 2022. Deep Learning for Text Style Transfer: A Survey. Computational Linguistics 48, 1 (March 2022), 155–205.

[23] Md Tawkat Islam et al Khondaker. 2024. DetoxLLM: A Framework for Detoxification with Explanations. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, 19112–19139.

[24] Satyapriya Krishna. 2023. On the Intersection of Self-Correction and Trust in Language Models. arXiv preprint arXiv:2311.02801 (2023).

[25] Léo et al Laugier. 2021. Civil Rephrases Of Toxic Texts With Self-Supervised Transformers. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, Online, 1442–1461.

[26] Beomseok et al Lee. 2024. XDetox: Text Detoxification with Token-Level Toxicity Explanations. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, 15215–15226.

[27] Alyssa et al Lees. 2022. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers (KDD ’22). Association for Computing Machinery, New York, NY, USA, 3197–3207.

[28] Varvara et al Logacheva. 2022. ParaDetox: Detoxification with Parallel Data. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 6804–6818.

[29] Do Xuan et al Long. 2024. Multi-expert Prompting Improves Reliability, Safety and Usefulness of Large Language Models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, 20370–20401.

[30] Scott M. et al Lundberg. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (Long Beach, California, USA) (NIPS’17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[31] Daniil et al Moskovskiy. 2024. LLMs to Replace Crowdsourcing For Parallel Data Creation? The Case of Text Detoxification. In Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, Miami, Florida, USA, 14361–14373.

[32] Cicero et al Nogueira dos Santos. 2018. Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 189–194.

[33] Skywork o1 Team. 2024. Skywork-o1 Open Series. https://huggingface.co/Skywork.

[34] OpenSource-O1. 2024. Open O1: A Model Matching Proprietary Power with Open-Source Innovation. Accessed: 2025-02-01.

[35] Ali et al Pesaranghader. 2023. Gpt-detox: An in-context learning-based paraphraser for text detoxification. In 2023 International Conference on Machine Learning and Applications (ICMLA). IEEE, 1528–1534.

[36] Mohammad Mahdi Abdollah et al Pour. 2023. COUNT: COntrastive UNlikelihood Text Style Transfer for Text Detoxification. In Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, 8658–8666.

[37] Emmanuelle-Anna Dietz Saldanha and Antonis C. Kakas. 2020. Cognitive Argumentation and the Suppression Task. CoRR abs/2002.10149 (2020). arXiv:2002.10149 https://arxiv.org/abs/2002.10149

[38] Anirudh et al Som. 2024. Demonstrations Are All You Need: Advancing Offensive Content Paraphrasing using In-Context Learning. In Findings of the Association for Computational Linguistics: ACL 2024, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12612–12627.

[39] Kaitao et al Song. 2020. Mpnet: Masked and permuted pre-training for language understanding. Advances in neural information processing systems 33 (2020), 16857–16867.

[40] Qwen Team. 2024. Qwen2.5: A Party of Foundation Models. https://qwenlm.github.io/blog/qwen2.5/

[41] Qwen Team. 2024. QwQ: Reflect Deeply on the Boundaries of the Unknown. https://qwenlm.github.io/blog/qwq-32b-preview/

[42] Josiane et al Van Dorpe. 2023. Unveiling Identity Biases in Toxicity Detection : A Game-Focused Dataset and Reactivity Analysis Approach. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track. Association for Computational Linguistics, Singapore, 263–274.

[43] Alex et al Warstadt. 2019. Neural Network Acceptability Judgments. Transactions of the Association for Computational Linguistics 7 (2019), 625–641.

[44] Jason et al Wei. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the 36th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.

[45] John et al Wieting. 2019. Beyond BLEU: Training Neural Machine Translation with Semantic Similarity. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 4344–4355.

[46] Rongwu et al Xu. 2024. 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. Association for Computational Linguistics, Miami, Florida, USA, 8341–8368.

[47] Chiyu et al Zhang. 2024. Distilling Text Style Transfer With Self-Explanation From LLMs. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop). Association for Computational Linguistics, Mexico City, Mexico, 200–211.

[48] Xuhui et al Zhou. 2021. Challenges in Automated Debiasing for Toxic Language Detection. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, Online, 3143–3155.

[49] Yuxiang et al Zhou. 2024. The Mystery of In-Context Learning: A Comprehensive Survey on Interpretation and Analysis. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, 14365–14378.

A Methodology

In this appendix we present the aggregation method for the Shapley values (Section A.1), and the different prompts used to generate the paraphrases (Section A.2).

A.1 Shapley Value Aggregation

In Section 3 we explained that five toxicity detectors are used to generate the aggregated Shapley values. To mitigate potential issues of robustness, bias, generalization, and false positives, especially in cases of implicit toxicity, we apply different aggregation methods based on each model's prediction performance. For comments where at least one model accurately predicts the class with a probability of 0.7 or higher, we aggregate the Shapley values by selecting only the tokens that are commonly identified as toxic across all models. In cases where all models incorrectly classify the comment, we aggregate the Shapley values of all models, as they at least identify toxic words that, in other contexts, could be considered negative. By considering all extreme cases, we account not only for clearly toxic words, but also for potentially toxic words that classifiers may misclassify due to performance limitations. This adds noise helps capture subtle or context-dependent toxic elements, challenging the LLMs' contextual understanding.

A.2 R^3-Detox Instructions

In the R^3-Detox framework, several prompts have been used: prompts to generate the reasoning for the ParaDetox, APPDIA, Parallel Detoxification, and Jigsaw Unintended Bias datasets, as well as the prompt used to generate the final non-toxic paraphrases. For all prompts, we use a structured chat template. Given the Self-Reflection's own instruction prompt, we introduce our task prompt as the user message, as it was experimentally observed that failing to include each model's system prompt caused the model to skip generating its reasoning tokens, thus producing a response as a standard instruction-following model without leveraging its self-reflection training. The instructions for generating the reasoning for the toxic and non-toxic sentences are the following:

```
Given a toxic sentence and its non-toxic paraphrase, your
task is to create a reasoning process for generating the
paraphrase. Avoid addressing the provided sentence as
"toxic sentence"; just call it the "provided sentence".
1. Examine the words or phrases in the provided sentence
that convey toxic behavior, and explain why they are
toxic within the given context. To support this analysis,
words that might indicate toxicity may be provided, but
avoid explicitly mentioning that these words were given.
2. Provide a concise explanation of why the sentence
exhibits toxic behavior based on your analysis.
3. Based on the toxic word analysis and the modifications
present in the paraphrase, analyze the changes required
to transform the sentence into the non-toxic paraphrase.
Begin by identifying the changes without directly stating
that the non-toxic paraphrase was provided, and later
explain how these changes remove the toxicity.

Be concise in your response. Here is the information to
analyse:
Toxic sentence: {sentence}
Relevant words: {toxic_words}
Label: {label}
Paraphrase: {paraphrase}
```

The instructions for generating the final paraphrase are as follows:

~~~
Provided a toxic sentence and words that may convey toxic
meaning, your task is to generate a non-toxic rephrased
version of the toxic sentence while maintaining the same
meaning and matching the target style/attribute as much
as possible. During your task, generate step-by step
reasoning following these instructions:
1. Examine the words or phrases in the provided sentence
that convey toxic behavior, and explain why they are
toxic within the given context. To support this analysis,
words that might indicate toxicity may be provided.
2. Provide a concise explanation of why the sentence
exhibits toxic behavior based on your
analysis.
3. Based on the toxic word analysis, analyze the changes
required to transform the sentence
into a non-toxic paraphrase. Begin by identifying the
changes needed and later explain how
these changes remove the toxicity.
The final output must be the following plain text:
```Final Reasoning: The reasoning process generated
Final Paraphrase: The final paraphrase generated```
**Example {str(idx+1)}:**
Toxic sentence: {sentence}
Possible toxic words: {toxic_words}
```Final Reasoning:"{reasoning}"
Final Paraphrase:"{paraphrase}"```
~~~

B Evaluation

In this appendix we introduce the metrics used to evaluate detoxification (Section B.1), the prompts used for evaluation with JudgeLLM (Section B.2), the guidelines (B.3) and annotation software (Section B.4).

B.1 Detoxification Metrics

In this section we provide a detailed explanation of the evaluation metrics introduced in Section 4.2, which are used to assess the quality of the generated non-toxic paraphrases. Each metric evaluates a different aspect of the detoxification process, including the preservation of meaning, fluency, and the reduction of toxicity. The following is a summary of the metrics:

- **Style Transfer Accuracy (STA)**: The percentage of non-toxic outputs identified by a style transfer model [28].
- **BERTScore**: We use the SimCSE [13] RoBERTa model to assess how well the model preserves the semantic meaning across tokens.
- **Content Preservation (SIM)**: The cosine similarity between the embeddings of the original toxic sentence and its paraphrase, computed using the model from Wieting [45].
- **Fluency (FL)**: The percentage of fluent sentences identified by a classifier trained on linguistic acceptability [43].
- **Joint Score (J)**: Quantifies the overall detoxification of the text as the product of the STA, SIM, and FL scores.
- **Toxicity Score (Tox)**: The probability that the text is toxic, as provided by the implementation of Hanu and Unitary team [17], which was used in Som [38] to compute the toxicity reduction.

B.2 JudgeLLM Evaluation System

During our research, JudgeLLM was used to generate evaluations for both the reasoning and the paraphrases. We designed two prompts: one for the "A vs. B" evaluation framework, used for reasoning evaluation, and another for the triplet tournament elimination framework, used for paraphrase evaluation. For the "A vs. B" evaluation framework prompt, we utilize the same prompt as for the generation of the reasoning as introduced in Section A.2. In the case of triplet tournament elimination evaluation framework, the prompt is the following:

```
Given a toxic sentence and words that may convey a toxic
meaning, your task is to generate a non-toxic rephrased
version of the sentence while maintaining the same
meaning and matching the target style or attribute as
closely as possible. A perfect non-toxic paraphrase is
one that perfectly preserves the meaning, is inoffensive,
and makes the least modifications to the original
sentence. Avoid generating a paraphrase that either
preserves the meaning but retains some offensiveness or
is inoffensive but significantly alters the original
meaning.
Toxic sentence: {sentence}
Possible toxic words: {toxic_words}
```

B.3 Manual Annotation Reasoning

For the manual annotation of the generated reasonings, we use the Argilla Annotation framework deployed on the Hugging Face Hub. As mentioned in Section 4, we implement a "A vs B" tournament based on the following guidelines:

```
In this dataset, you will be presented with a prompt and
two reasoning responses. Your task is to select the
reasoning response that best fits the given prompt. If
both reasoning responses are equally good or bad, select
0. 1 corresponds to Model A and 2 corresponds to Model B.
Please rank the reasoning response based on the following
criteria:
1. Task Completion: Does the reasoning correctly tackle
the task given in the prompt? For prompts involving toxic
sentence analysis, does the reasoning tackle the
following goals: (a) analyze words that may convey toxic
meaning, (b) provide a verdict on whether it is toxic,
and (c) if toxic, reason correctly on how to obtain the
final paraphrase.
2. Coherence: Is the reasoning response coherent and
logically sound? Does it contain contradictions or broken
logical flow or misleading statements?
3. Fluency: Is the reasoning response fluent and easy to
understand? Does it contain any grammatical errors?
4. Consistency: In the cases where a toxic sentence is
analyzed, does the reasoning stay consistent with the
provided paraphrase?
5. Plausibility: Is the reasoning response plausible and
realistic? Is the contextual analysis provided by the
models believable given the task?
6. Creativity: Does the reasoning response show
creativity and originality?
```

B.4 Manual Annotation Paraphrase

Like the reasoning process, manual annotation of generated paraphrases uses the Argilla framework on Hugging Face. As described in Section 4, annotators rank the best paraphrase following these guidelines:

```
You will be presented with a prompt and three model
responses. Your task is to select the non-toxic
paraphrase that best fits the criteria. Only one response
can be ranked as the best. Rank the paraphrases based on
the following criteria, where the best value is a 5 and
the worst is a 1:
5. A perfect meaning-preserving, inoffensive paraphrase
that not only preserves the meaning but also makes the
least modifications to the original sentence.
4. A perfect meaning-preserving, inoffensive paraphrase
but with significant modifications to the original
sentence.
3. A paraphrase that is inoffensive but somewhat distinct
in meaning from the original.
2. A meaning-preserving paraphrase that is somewhat
similar in offensiveness to the original.
1. A paraphrase that is very different in meaning and not
less offensive than the original.
```

Received 5 September 2025; accepted 22 October 2025