Evaluating Text Style Transfer: A Nine-language Benchmark for Text Detoxification

Vitaly Protasov[1], **Nikolay Babakov**[2], **Daryna Dementieva**[3,4], **Alexander Panchenko**[5,1]
[1]AIRI [2]Universidade de Santiago de Compostela [3]Technical University of Munich (TUM)
[4]Munich Center for Machine Learning (MCML) [5]Skoltech
protasov@airi.net, nikolay.babakov@usc.es, daryna.dementieva@tum.de, a.panchenko@skol.tech

Abstract

Despite notable advances in large language models (LLMs), reliable evaluation of text generation tasks such as text style transfer (TST) remains an open challenge. Existing research has shown that automatic metrics often correlate poorly with human judgments (Dementieva et al., 2024; Pauli et al., 2025), limiting our ability to assess model performance accurately. Furthermore, most prior work has focused primarily on English, while the evaluation of multilingual TST systems, particularly for text detoxification, remains largely underexplored. In this paper, we present the first comprehensive multilingual benchmarking study of evaluation metrics for **text detoxification** evaluation across *nine languages*: Arabic, Amharic, Chinese, English, German, Hindi, Russian, Spanish, and Ukrainian. Drawing inspiration from machine translation evaluation, we compare neural-based automatic metrics with LLM-as-a-judge approaches together with experiments on task-specific fine-tuned models. Our analysis reveals that the proposed metrics achieve significantly higher correlation with human judgments compared to baseline approaches. We also provide actionable insights and practical guidelines for building robust and reliable multilingual evaluation pipelines for text detoxification and related TST tasks.

Keywords: multilingual evaluation, text style transfer, text detoxification, large language models

1. Introduction

Evaluation of text generation tasks remains a long-standing challenge in natural language processing (NLP), as suitable metrics must capture both task-specific objectives and linguistic variation across languages. The difficulty of designing such metrics increases with the number of languages considered, since stylistic, morphological, and pragmatic differences demand diverse evaluation approaches. Over the past decade, the field has seen continuous evolution of evaluation methodologies: from character-based measures such as chrF (Popovic, 2015), to lexical and semantic metrics like ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2019), and more recently to neural-based models such as COMET (Rei et al., 2020). With the emergence of large language models (LLMs), new evaluation paradigms—particularly *LLM-as-a-judge* setups—have demonstrated strong potential for approximating human judgments (Bavaresco et al., 2025).

In this study, we focus on a specific text style transfer (TST) task: **text detoxification**. The goal of this task is to remove toxic or offensive content from text while preserving the original meaning and fluency. Beyond its research value, text detoxification is of practical importance for a wide range of real-world applications, including online moderation, dialogue systems, and social media content filtering and post-processing. For instance, consider a video streaming service which can offer a dynamic rewriting of abusive words enabling a child-friendly watching. Another important use-case is to implement guards against an eventual generation of abusive words by a neural dialogue system for client support.

Despite its relevance, the task remains underexplored in both data and evaluation methodology. To date, only two publicly available datasets exist: TextDetoxEval (Dementieva et al., 2024), covering nine languages, and DialogueEvaluation-2022 (Dementieva et al., 2022), which focuses solely on Russian. Moreover, evaluation practices in this area are often inconsistent, with most prior studies relying on generic metrics rather than task-specific or multilingual approaches.

To address these gaps, we perform a comprehensive experimental study of evaluation metrics for the text detoxification task across **nine languages**: Arabic, Amharic, Chinese, English, German, Hindi, Russian, Spanish, Ukrainian. Our goal is to analyze the behavior, limitations, and cross-lingual robustness of different metric families, including both automatic neural-based metrics and LLM-as-a-judge approaches. In addition, we perform fine-tuning experiments with one of open-source LLMs on annotated data. We aim to identify metrics that align more closely with human judgments and to provide guidance for future evaluation of detoxification and related TST systems. The main contributions of this work are:

1. We conduct an extensive multilingual evaluation study for the text detoxification task, covering all publicly available datasets for this task.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4560–4574
11–16 May 2026.

2. We experiment with a diverse set of evaluation metrics, proposing several improved metric configurations tailored to text detoxification.
3. We compare all approaches against existing automated and LLM-as-a-judge setups, highlighting their respective strengths and weaknesses across languages.
4. We fine-tune open-source LLMs on annotated detoxification data and evaluate their suitability as automatic judges.

To facilitate future research and reproducibility, we make our evaluation setup, code, and results publicly available.[1] Details of all resources, corresponding links, and their licenses are presented in Appendix A.

2. Related Work

2.1. Automatic Evaluation for TST and Text Detoxification

The task of text style transfer (TST) has been studied across a variety of domains. This includes sentiment transfer, such as converting between positive and negative reviews (Li et al., 2018); formality transfer (Rao and Tetreault, 2018; Briakou et al., 2021), which focuses on transforming informal texts into formal ones; and stylistic rewriting, exemplified by the Bible style transfer task (Carlson et al., 2018), which leverages translations from different historical periods. Additionally, the biased-to-neutral Wikipedia corpus (Pryzant et al., 2020) makes use of editorial revisions to reduce bias. In the domain of text detoxification, the task has been addressed already with several solutions using both unsupervised approaches (Nogueira dos Santos et al., 2018; Dale et al., 2021; Hallinan et al., 2023) and supervised methods, supported by parallel data (Logacheva et al., 2022; Mukherjee et al., 2023; Rykov et al., 2024). Across various domains, the evaluation of text style transfer systems has traditionally relied on three core criteria:

- *Style Accuracy* (**STA**): the proportion of outputs correctly classified in a new style by a style classifier.
- *Content Preservation* (**SIM**): the extent to which the key semantic content from the original input is retained.
- *Fluency* (**FL**): whether the generated output maintains natural fluency or, at minimum, does not degrade the fluency of the original.

Considerable efforts have been made to develop more robust evaluation metrics for text style transfer (Bavaresco et al., 2024; Pauli et al., 2025) and, in particular, for text detoxification (Dementieva et al., 2023), a universally accepted automatic evaluation framework that strongly correlates with human judgment has yet to be established. Thus, as a recent state-of-the-art evaluation setup, we adopt setup from Dementieva et al. (2024) where: (i) **STA** is defined as the probability assigned by a pre-trained XLM-R (Conneau et al., 2020) toxicity classifier indicating that the output text belongs to the neutral (non-toxic) class; (ii) **SIM** is calculated as the cosine similarity between LaBSE (Feng et al., 2022) representations of toxic and detoxified texts; (iii) **FL**: measured as a proxy using the ChrF (Popovic, 2015), which compares the generated detoxified text to human-written references. These all three parameters are then combined into joint metric **J**:

$$\mathbf{J} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{STA}(y_i) \cdot \mathbf{SIM}(x_i, y_i) \cdot \mathbf{chrF}(x_i, y_i), \quad (1)$$

where $\mathbf{STA}(y_i)$, $\mathbf{SIM}(x_i, y_i)$, $\mathbf{chrF}(x_i, y_i) \in [0, 1]$ for each text detoxification output y_i. Here x_i means detoxified golden texts and y_i is detoxified generated texts.

Limitations of Current Metrics Current evaluation metrics for detoxification are hindered by their limited and often superficial use of human reference texts. Thus, only *fluency*, assessed via ChrF, explicitly leverages references. However, ChrF suffers from a core limitation: it evaluates surface-level n-gram overlap with the reference, ignoring the semantic relationship between the system output and the original toxic input. This reliance on lexical similarity renders the metric both *overconstrained*—discouraging variation—and *underconstrained*—failing to ensure semantic preservation. These issues underscore the need for evaluation approaches that more effectively integrate both source and reference relationships.

2.2. LLMs as a Judge

The emergence of large language models (LLMs) has introduced a new paradigm in evaluation, where LLMs themselves are used as judges for NLP tasks (Li et al., 2024). This approach has been explored by Pauli et al. (2025) across 20 NLP tasks, including text style transfer (TST). In the domain of abusive language, LLMs have been employed to assess the relevance and appropriateness of counter-speech responses to hate speech (Jones et al., 2024; Bonaldi et al., 2024). TST can also be framed as a paraphrasing task, where LLMs-as-judges have shown potential (Lemesle et al., 2025). While LLM-based evaluation is not without limitations, it offers a promising and adaptable solution—particularly for multilingual contexts.

[1] https://github.com/textdetox/eval-of-detox-eval

3. Evaluation Datasets

Our experiments rely on two available datasets for text detoxification task: TextDetoxEval (Dementieva et al., 2024) and DialogueEvaluation-2022 (Dementieva et al., 2022), with most of the focus on the first one due to its multilinguality.

3.1. TextDetoxEval

TextDetoxEval[2,3] is a multilingual dataset released as part of the CLEF Text Detoxification Shared Task. It includes manual assessments of 20 detoxification systems across 9 languages: Amharic, Arabic, Chinese, English, German, Hindi, Russian, Spanish, and Ukrainian. The systems covered a range of modeling strategies, from unsupervised to fine-tuned and LLM-based prompting approaches (Peng et al., 2024; Luo et al., 2024; Protasov, 2024). For each language, 100 toxic sentences were selected, resulting in 900 inputs overall. Each of the 20 participant generated detoxified versions, producing **16,600** input–output pairs in total with corresponding human scores. Native speakers annotated the data on Toloka.ai[4] according to three criteria:

- **Fluency:** grammaticality and readability rated as *yes*, *partially*, or *no*.
- **Content Similarity:** whether the detoxified text preserves the meaning of the original (binary judgment).
- **Style Transfer Accuracy:** which text is more toxic (*original*, *detoxified*, or *neither*), randomized to avoid bias.

3.2. DialogueEvaluation-2022

DialogueEvaluation-2022[5] is a Russian-language text detoxification dataset released as part of the first shared task on detoxification. The dataset contains toxic sentences collected from social media platforms. It includes outputs from **15 participating solutions**, each generating detoxified rewrites for a shared test set of **875 toxic sentences**.

Data collection followed a three-stage crowdsourcing procedure on Toloka. First, annotators rewrote toxic sentences into fluent, non-toxic paraphrases while preserving meaning. Second, independent annotators verified semantic equivalence between the original and rewritten texts. Third, additional annotators checked that toxicity had been successfully removed. Only paraphrases that passed both verification stages with at least 90% agreement were retained. Each text pair was further evaluated along three dimensions consistent with the TextDetoxEval *fluency*, *content preservation*, and *style transfer quality*. This dataset provides a valuable complementary resource for benchmarking Russian-language detoxification and cross-lingual metric evaluation.

Although DialogueEvaluation-2022 offers useful monolingual insights, we focus the main analysis of this paper on the multilingual TextDetoxEval dataset to ensure consistency across languages and to maintain a unified evaluation setup. The detailed results for DialogueEvaluation-2022 are provided in Appendix C.

4. Evaluation Methodology

In this section, we first aim to describe how evaluation in text detoxification task has been performed in previous works (Dementieva et al., 2022, 2024), and second, to propose ways to improve it by exploring alternative approaches or improving existing ones. Afterwards, we conduct experiments based on the described approaches to benchmark them and assess whether our proposed improvements are effective.

4.1. Fluency

As noted in Section 2.1, current fluency evaluation in text detoxification typically relies on ChrF scores calculated between system outputs and human references. However, this approach ignores the original toxic input. As a result, it may favor outputs that look similar to the reference but do not fully preserve the meaning or context of the input sentence. In detoxification tasks, this becomes a problem because the goal is not only to make the text fluent and non-toxic, but also to preserve the meaning and context of the original text.

To address this limitation, we aim to improve fluency evaluation by considering more advanced neural-based models, namely **COMET** (Rei et al., 2020) and **XCOMET** (Guerreiro et al., 2023).

Unlike traditional metrics that rely solely on lexical overlap with a reference, COMET-based models use pretrained encoders to model semantic relationships between the *input*, *system output*, and *reference*. This triplet-based setup allows them to jointly assess whether the generated text maintains the original meaning and intent. In machine translation, COMET models demonstrated significantly higher correlation with human judgments compared to n-gram metrics, as they better capture meaning, syntax, and fluency beyond token overlap.

In the context of text detoxification, we hypothesize that COMET models are suitable, as they

[2] https://hf.co/datasets/textdetox/detoxification_pairwise_style_evaluation

[3] https://hf.co/datasets/textdetox/detoxification_pairwise_content_evaluation

[4] https://toloka.ai

[5] https://hf.co/datasets/textdetox/humaneval_textdetox_ru

should better capture the trade-off between fluency, content preservation, and toxicity reduction, which are key aspects that simpler lexical metrics often fail to account for. In our experiments, we consider 4 COMET variants:

1. **Unbabel/wmt22-comet-da** (Rei et al., 2022): A regression-based model trained on direct assessment data from WMT22, representing a standard reference-aware evaluation setting.
2. **Unbabel/XCOMET-XL** (Guerreiro et al., 2023): A multilingual extension of COMET with a 3.5B-parameter encoder, improving correlation with human judgments across languages.
3. **Unbabel/XCOMET-XXL** (Guerreiro et al., 2023): A 10.7B-parameter variant achieving state-of-the-art performance on WMT22 evaluation benchmarks.
4. **myyycroft/XCOMET-lite** (Larionov et al., 2024): A compressed and quantized model retaining over 95% of XXL's performance while reducing computational overhead by 60%, enabling scalable multilingual evaluation.

4.2. Content similarity

The content similarity score measures how well the generated text preserves the main semantic information of the original input. This metric penalizes outputs that lose important details or change the original meaning during the detoxification process. As noted in Section 2.1, previously content similarity was computed using cosine similarity between the embedding representations of source toxic texts and their detoxified versions.

This approach completely ignores available reference texts, which creates two significant problems. First, metrics that only compare input to output perform poorly when the generated text requires substantial rewording—exactly what detoxification demands (Shen et al., 2022). Second, without references, the metric cannot determine whether semantic content is genuinely preserved when lexical choices differ substantially from the input (Shen et al., 2022).

To address these limitations, we propose an improved content similarity measurement metric that jointly considers both the input–output and output–reference relationships:

$$\mathrm{c}_{\mathrm{sim}} = w_{\mathrm{i,g}} \cdot \cos_{\mathrm{sim}}(v_{\mathrm{i}}, v_{\mathrm{g}}) + w_{\mathrm{g,r}} \cdot \cos_{\mathrm{sim}}(v_{\mathrm{g}}, v_{\mathrm{r}}). \quad (2)$$

where v_{i}, v_{g}, and v_{r} denote the embedding representations of the **input**, **generated**, and **reference** texts, respectively. The weights $w_{\mathrm{i,g}}$ and $w_{\mathrm{g,r}}$ control the contribution of input–output and output–reference similarity, with $w_{\mathrm{i,g}} + w_{\mathrm{g,r}} = 1$. This approach aims to balance two complementary aspects of content preservation: semantic consistency with source inputs and alignment with human-produced references, which together provide a more complete estimate of content preservation.

4.3. Style transfer performance: toxicity

Toxicity measurement reflects how well a system transforms toxic inputs into neutral or polite outputs. In previous works, toxicity was measured using the probability of the generated text being classified as *neutral* by a binary toxicity classifier. However, relying solely on this probability introduces several limitations: (i) high dependence on the biases and calibration of the specific classifier; (ii) toxicity of the original input is not taken into account, making it impossible to measure relative improvement; and (iii) reference texts, which represent human-level detoxification quality, are ignored entirely.

To provide a more robust assessment, we propose evaluating style transfer through comparative probability analysis across three text variants:

- Input (toxic) text: $P_{\mathrm{neutral}}(t_{\mathrm{i}})$,
- Generated text: $P_{\mathrm{neutral}}(t_{\mathrm{g}})$,
- Reference (neutral) text: $P_{\mathrm{neutral}}(t_{\mathrm{r}})$.

This setup enables the evaluation to capture the relative change in toxicity rather than relying on an absolute classifier score. In other words, instead of only checking whether an output is predicted as neutral, we measure how well it has improved compared to its toxic input, and whether it approaches the quality of a human-written neutral reference. Such relative comparisons are less sensitive to classifier calibration and better reflect the underlying goal of detoxification — to reduce toxicity while preserving semantic content.

To ensure consistent behavior across diverse systems, we introduce two stabilization rules:

1. **Penalization:** if the generated text is more toxic than the input ($P_{\mathrm{neutral}}(t_{\mathrm{g}}) < P_{\mathrm{neutral}}(t_{\mathrm{i}})$), the score is set to zero: $Score(t_{\mathrm{g}}) = 0$;
2. **Rewarding:** if the generated text achieves neutrality equal to or greater than the reference ($P_{\mathrm{neutral}}(t_{\mathrm{g}}) \geq P_{\mathrm{neutral}}(t_{\mathrm{r}})$), the score is set to one: $Score(t_{\mathrm{g}}) = 1$.

This triplet-based design makes the toxicity metric more robust as it mirrors the logic used in COMET and content similarity metrics, where relationships between input, output, and reference are jointly modeled. As a result, the metric better reflects meaningful improvements in detoxification quality and mitigates dependence on specific classifiers.

4.4. LLMs as judges

To complement proposed automatic metrics and assess their consistency with human evaluations, we

additionally employ LLMs as automatic evaluators to compare how well different metric-based families (lexical, neural, and LLM-based) align with human annotations in the context of text detoxification.

4.5. Fine-tuning of LLMs

We also investigate whether task-specific fine-tuned models can provide evaluations that better align with human judgments than their general-purpose counterparts. To this end, we conduct fine-tuning experiments on Llama-3.1-8B[6].

5. Results

5.1. Experimental Setup

This section presents the results of our experiments following the methodology outlined in Section 4. We first compare the performance of previously adopted evaluation metrics used in the target shared task (Dementieva et al., 2024), which introduced the multilingual TEXTDETOXEVAL dataset. We then analyze the performance of our proposed metrics and examine how LLMs perform as automatic evaluators. Finally, we report results from LLM fine-tuning experiments designed to establish comprehensive benchmarks across multiple aspects of the text detoxification task. The results obtained on the DialogueEvaluation-2022 dataset in Russian (Dementieva et al., 2022) are reported in Appendix C.

To ensure a fair comparison between LLM-as-a-judge and the fine-tuning experiments, we split the TEXTDETOXEVAL dataset by dividing the 20 participating systems into training and test sets, where outputs from 14 systems are used for training, while outputs from the remaining 6 systems are reserved for testing. This participant-based split ensures that fine-tuned models are evaluated on outputs from previously unseen systems. All results reported for both LLM-as-a-judge and fine-tuning experiments are computed on this held-out test set.

To assess the quality and reliability of different evaluation approaches, we measure the agreement between automatic metrics and human judgments using Spearman's rank correlation coefficient (ρ). This non-parametric measure is particularly well-suited for our study as it captures monotonic relationships without assuming linearity, making it robust to the varying scales and distributions inherent in different metric families. For each evaluation dimension—*fluency*, *content similarity*, and *toxicity*—we compute correlations between metric scores and human annotations across all system outputs. We report correlations separately for each language to account for linguistic variation and to identify cross-lingual patterns in metric behavior.

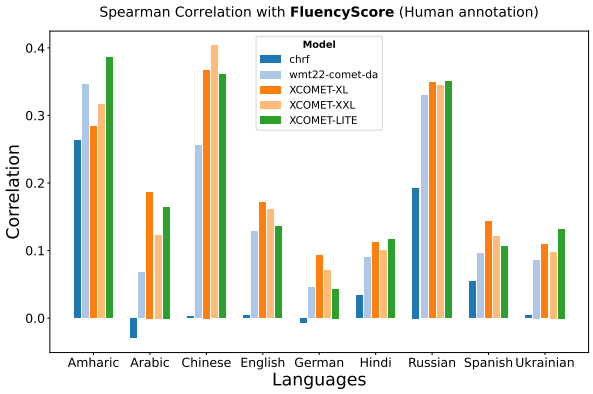

Figure 1: TEXTDETOXEVAL: Correlation of fluency measurement approaches with human-annotated fluency scores.

5.2. Fluency Evaluation Results

Figure 1 presents Spearman correlations between automatic fluency metrics and human judgments across all languages and models. ChrF consistently underperforms, showing near-zero correlations in Arabic, German, Chinese, English, and Ukrainian, which can be attributed to its reliance on character n-gram overlap with references, penalizing semantically equivalent paraphrases that use different word choices or phrasing, a common requirement in detoxification where toxic phrases must be substantially rewritten. In contrast, COMET-based models demonstrate positive correlations across languages, with **XCOMET-XXL** and **XCOMET-lite** achieving the strongest performance in Amharic, Chinese, and Russian.

The magnitude of correlations varies considerably, ranging from weak in German and Hindi to moderate in Amharic and Russian, suggesting that fluency evaluation difficulty is language-dependent. This variation may stem from differences in training data availability for these languages in XCOMET's multilingual pre-training, as well as varying degrees of lexical diversity in the reference translations across languages. Notably, a quantized model **XCOMET-lite** maintains competitive performance with **XCOMET-XXL**, making it an optimal choice for future detoxification competitions and production systems that require reliable evaluation without extensive computational resources.

5.3. Content Similarity Evaluation Results

Figure 2 presents Spearman correlations between automatic content similarity metrics and human judgments across all languages. We evaluate three

[6] https://hf.co/meta-llama/Llama-3.1-8B

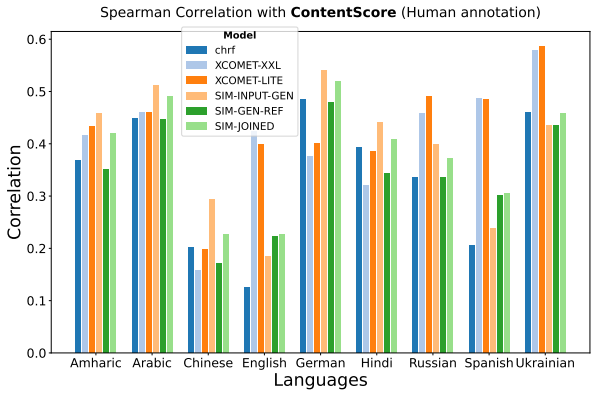

Figure 2: TEXTDETOXEVAL: Correlation of content similarity measurement approaches with human-annotated content preservation scores.

embedding-based configurations: (i) **SIM-INPUT-GEN**, cosine similarity between the toxic input and generated output, serving as the baseline approach from prior work (Dementieva et al., 2024); (ii) **SIM-GEN-REF**, similarity between the generated output and human reference; and (iii) **SIM-JOINED**, our proposed weighted combination defined in Equation 2, where $w_{i,g} = 0.4$ and $w_{g,r} = 0.6$. We assign higher weight to the generated-reference similarity because human references represent high-quality detoxification examples that balance content preservation with appropriate rewording, while input-output similarity may be misleading when toxic phrasing requires substantial changes.

Surprisingly, the results reveal that the baseline **SIM-INPUT-GEN** achieves the highest correlations with human judgments in five languages (Amharic, Arabic, Chinese, German, Hindi). This unexpected finding suggests that human annotators may prioritize preserving the original text's meaning over adhering to the reference paraphrasing style when assessing content preservation. However, such a tendency can be detrimental to annotation consistency, as it may favor outputs that closely mirror the source text and penalize significant paraphrases that achieve equivalent meaning through different lexical or syntactic choices. In languages where detoxification requires substantial rewording to remove offensive content (Arabic, German), measuring similarity only to toxic inputs provides limited signal about whether appropriate semantic transfer has occurred.

Among neural models, **XCOMET-LITE** and **XCOMET-XXL** demonstrate strong and consistent performance across all languages. Notably, ChrF shows surprisingly competitive performance in Arabic, German, and Ukrainian, suggesting that character n-gram overlap can approximate content preservation when detoxification strategies involve minor lexical substitutions.

According to our results, we recommend prioritizing COMET-based models for evaluating content similarity in detoxification tasks. Unlike embedding-based metrics that capture only surface-level similarity to the input, XCOMET models jointly consider the relationships among the input, output, and reference texts. This joint modeling yields more stable and robust evaluations that are more resistant to superficial lexical changes. Moreover, the direct consideration of reference texts in XCOMET aligns more closely with the detoxification objective, producing fluent, neutral outputs that preserve meaning without reproducing toxic expressions.

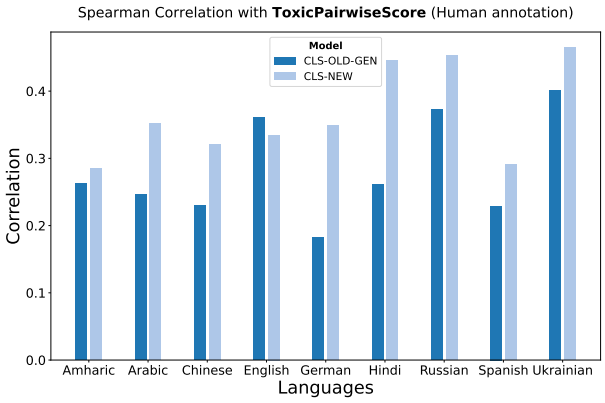

Figure 3: TEXTDETOXEVAL: Correlation of toxicity measurement approaches with target pairwise toxic human annotated scores.

5.4. Toxicity Evaluation Results

As described in Section 4.3, the effectiveness of style transfer in detoxification is typically assessed using a pretrained toxicity classifier, which estimates the probability that a generated text belongs to the *non-toxic* class.

Here we aim to compare two approaches: (i) **CLS-OLD-GEN**, which measures the predicted probability of the generated text being non-toxic; and (ii) **CLS-NEW**, a joint probability–based approach introduced in Section 4.3, which integrates toxicity signals from the input, generated output, and reference texts. Both approaches employ the same model[7] as prior work.

Figure 3 presents the correlation of these approaches with human-annotated toxicity scores. **CLS-NEW** achieves the highest correlation across all languages except English, where **CLS-OLD-GEN** performs slightly better. Strong cross-lingual performance of **CLS-NEW** highlights the advantage of incorporating contextual information from input–output–reference triplets. By jointly considering them, **CLS-NEW** provides more stable toxicity estimates, reducing sensitivity to minor lexical

[7] https://hf.co/textdetox/xlmr-large-toxicity-classifier

changes and enabling fairer, context-sensitive evaluation in the detoxification task.

5.5. Evaluation of Joined Metric (J)

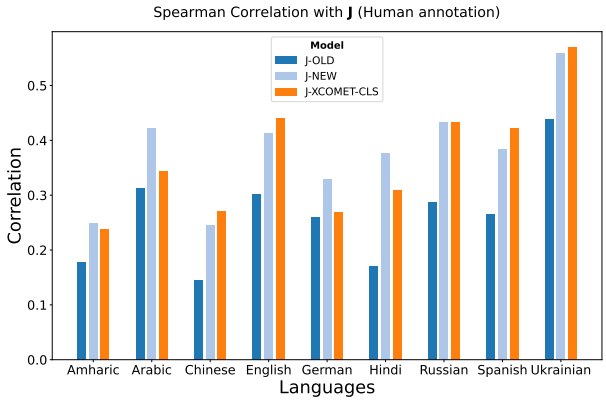

Figure 4: TEXTDETOXEVAL: Correlation final scores with target joined scores from human annotation.

Following the original formulation of the joint detoxification score presented in Equation 1, we evaluate how well different metric combinations correlate with human judgments. Figure 4 presents results for three composite evaluation approaches: (i) **J-OLD**, the baseline approach combining ChrF, SIM-INPUT-GEN, and CLS-OLD-GEN; (ii) **J-NEW**, our proposed combination of XCOMET-LITE, SIM-JOINED, and CLS-NEW; and (iii) **J-XCOMET-CLS**, a simplified variant using only XCOMET-LITE and CLS-NEW.

The rationale for considering J-XCOMET-CLS is to leverage XCOMET-LITE's ability to assess both fluency and content similarity from input–output–reference triplets, eliminating the need for a separate content similarity metric.

The results show that **J-OLD** achieves the lowest correlations across all languages, highlighting the limitations of the baseline approach. Our proposed **J-NEW** method demonstrates the highest correlations in 5 languages (Amharic, Arabic, German, Hindi, Russian). Interestingly, **J-XCOMET-CLS** performs best in 4 languages (Chinese, English, Spanish, Ukrainian), slightly outperforming **J-NEW**, but showing lower correlations in Arabic, German, and Hindi. These findings suggest that XCOMET-LITE can effectively capture both fluency and content similarity in the text detoxification task.

5.6. Evaluation using LLMs as Judges

In addition to aforementioned methods, we also explore the use of LLMs as evaluators. We consider a diverse set of models: DeepSeek-R1-Distill-Qwen-32B[8], DeepSeek-V3-0324[9], LLaMA 3.3-70B-Instruct[10], CompassJudger-1-32B-Instruct (Cao et al., 2024), and GPT-4.1 variants (*nano* and *mini*). These models represent a diverse range of architectures and scales, allowing us to examine whether larger or more instruction-tuned models offer improved alignment with human judgments across detoxification evaluation dimensions. All final prompts are provided in Appendix B.

Fluency Figure 5 depicts correlation results with human annotations, comparing XCOMET-LITE against six LLM-as-a-judge approaches (Llama-3.1-8b-finetuned is considered in the next subsection). XCOMET-LITE significantly outperforms LLM baselines only in Amharic and Chinese. In contrast, LLaMA 3.3-70B-Instruct-Turbo achieves the highest correlations in Arabic, Hindi, Russian, and Ukrainian, substantially surpassing other models.

Content similarity Figure 6 compares the proposed SIM-JOINED metric with six LLM-as-a-judge approaches. SIM-JOINED achieves the highest correlations in Amharic, Arabic, Chinese, German, Hindi, and Ukrainian, outperforming LLM-based evaluation in most cases. This is in contrast to fluency evaluation, where XCOMET-LITE performed worse than LLMs across most languages. This difference can be explained by the nature of the tasks: content similarity is a clear semantic matching problem, where embedding-based metrics perform well by measuring the proximity of text representations. In contrast, fluency requires assessing grammaticality, naturalness, and stylistic appropriateness factors that XCOMET-LITE struggles to capture as effectively as LLMs in this detoxification task.

Toxicity In Figure 7 the CLS-NEW model outperforms all LLMs only in Chinese and shows poor correlations in other languages. Notably, GPT-4.1-mini demonstrates strong performance across all languages; DeepSeek-R1-Distill-Qwen-32B achieves the highest correlations in Russian, significantly outperforming all other models, and demonstrates one of the best performance in Arabic and Spanish.

5.7. Fine-tuning Experiments

We perform fine-tuning of Llama-3.1-8B using the Low-Rank Adaptation (LoRA) (Hu et al., 2022) method to efficiently adapt the base models while minimizing computational overhead. The main model weights are loaded in 4-bit quantized format, enabling faster training and reduced memory usage without significant loss in performance. The

[8] https://hf.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
[9] https://hf.co/deepseek-ai/DeepSeek-V3-0324
[10] https://hf.co/meta-llama/Llama-3.3-70B-Instruct

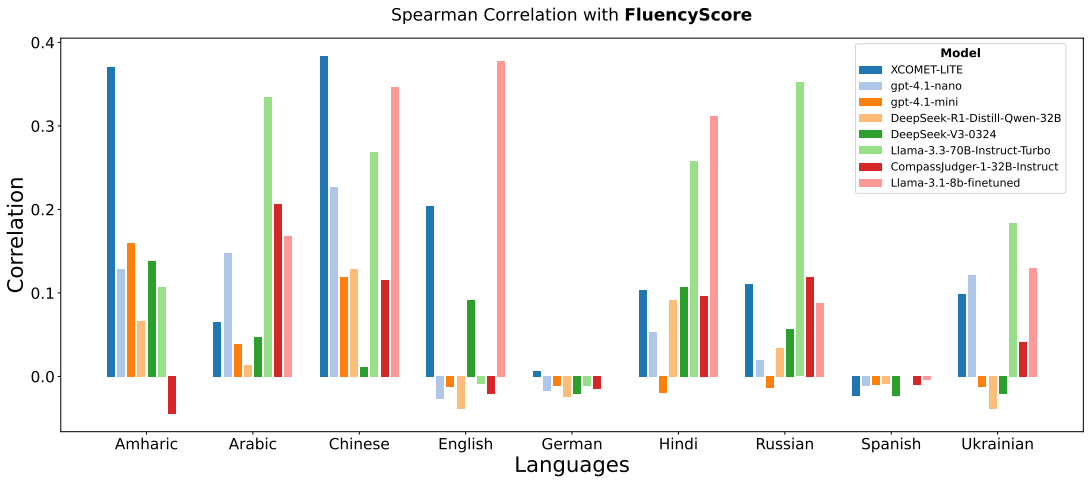

Figure 5: TEXTDETOXEVAL: Comparison between **XCOMET-LITE** and different LLMs on the fluency scores from human annotation.

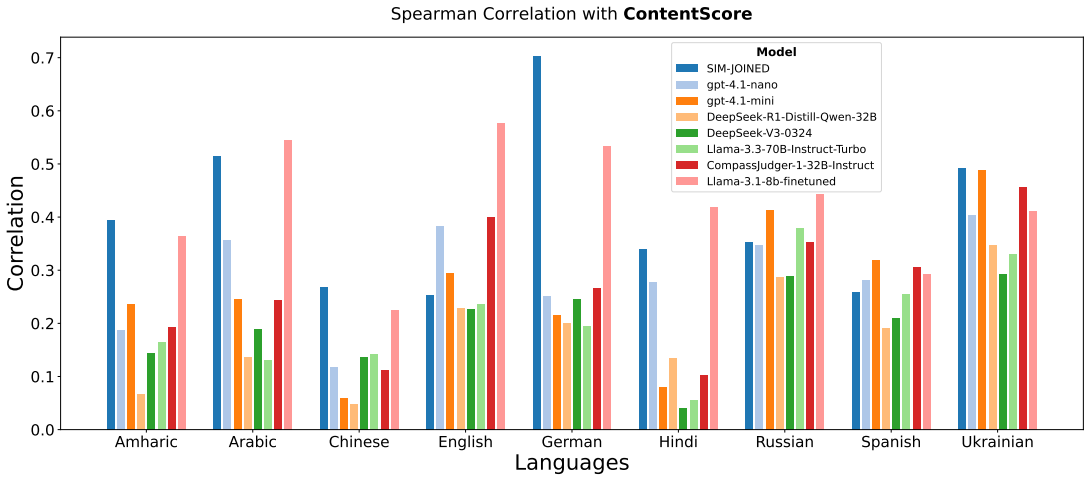

Figure 6: TEXTDETOXEVAL: Comparison between **SIM-JOINED** and different LLMs on the content similarity scores from human annotation.

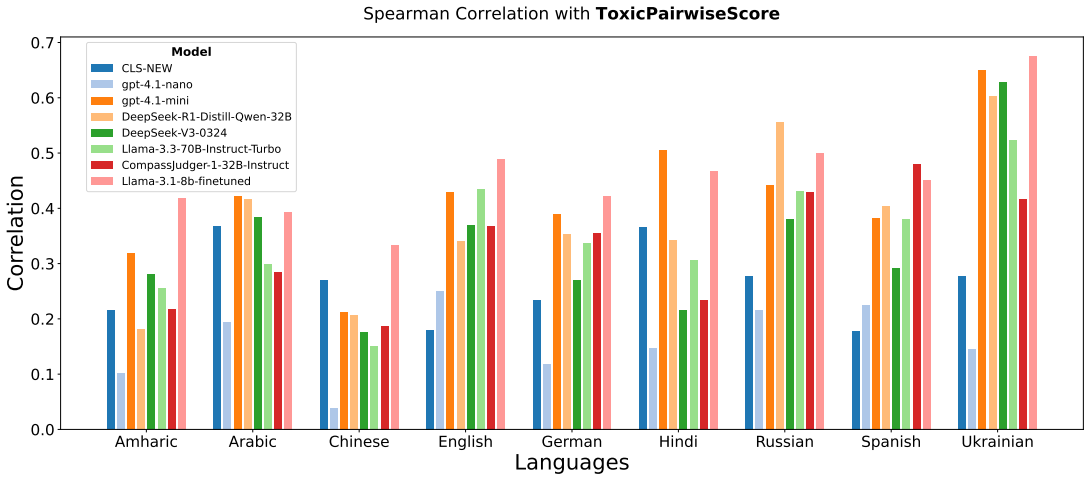

Figure 7: TEXTDETOXEVAL: Comparison between **CLS-NEW** and different LLMs on the toxicity classification scores from human annotation.

LoRA configuration utilizes the following hyperparameters: rank $r = 8$, $\alpha = 16$, and a dropout rate of 0.1, with adaptation applied to all linear layers.

The fine-tuning process was conducted for 2 epochs over the training data. For each task (i.e. content similarity, style transfer, or fluence) we fine-tuned standalone LORA. The resulting tuned LLMs are available in our HuggingFace repository.[11,12]

Figures 5, 6, and 7 present the fine-tuning results for Llama-3.1-8B across all evaluation dimensions. For fluency evaluation, the fine-tuned model achieves the highest correlations in English and demonstrates competitive performance in Chinese, while showing notably lower correlations in other languages. This pattern likely reflects the composition of Llama's pretraining data, where English is substantially more represented, enabling better adaptation to fluency assessment tasks in that language.

For content similarity evaluation, the fine-tuned model demonstrates strong and consistent performance across nearly all languages, achieving the highest or near-highest correlations except in Spanish and Ukrainian. Finally, for toxicity evaluation, the fine-tuned model shows stable performance across all languages, achieving the highest correlations in Amharic, Chinese, English, German, and Ukrainian.

In summary, the findings suggest that each evaluation task benefits differently from fine-tuning, with fluency being most sensitive to language coverage, while content and toxicity assessments exhibit stronger cross-lingual robustness.

6. Conclusion

We presented the first large-scale multilingual evaluation of text detoxification on the TextDetoxEval dataset, covering all core dimensions: fluency, content similarity, and toxicity reduction. To improve evaluation, we introduced XCOMET-based models for fluency, a triplet-based SIM-JOINED metric for content similarity that leverages input–output–reference texts, along with CLS-NEW, a refined toxicity metric that jointly considers probability distributions also across all text triplets.

We thoroughly analyzed different evaluation metrics and LLM configurations, including LLM-as-a-judge experiments and task-specific fine-tuning, to identify approaches that align most closely with human judgments across diverse languages.

We hope the benchmark results, improved metrics, and insights into evaluation approaches presented in this study facilitate future research on multilingual text detoxification and related style transfer tasks.

[11] https://hf.co/textdetox/Llama-pairwise-toxicity-evaluator

[12] https://hf.co/textdetox/Llama-pairwise-content-evaluator

7. Limitations

The experiments in this paper are limited by the number of languages considered. Currently, our conclusions are based solely on experiments involving 9 languages: Amharic, Arabic, Chinese, English, German, Hindi, Russian, Spanish, and Ukrainian. As a result, our findings may be biased toward these languages due to the lack of available data for others. The generalization of our proposed metrics to additional languages, particularly low-resource languages outside these families, remains an open question requiring further investigation.

Our evaluation relies on the TextDetoxEval and DialogueEvaluation-2022 datasets, which contain outputs from 20 and 15 detoxification systems, respectively. While these datasets represent diverse modeling approaches, the limited number of systems and the specific characteristics of their outputs may influence metric correlations. Future work should validate our findings on larger and more diverse collections of detoxification systems across additional domains.

Proposed XCOMET-based fluency metrics and fine-tuned LLM evaluators depend on the quality and coverage of their pre-training data. Languages with limited representation in these training corpora may receive less reliable evaluations. The observed performance variations across languages particularly the weaker correlations in German and Hindi suggest that data availability during model pre-training directly impacts evaluation quality.

8. Ethical Considerations

Our work on evaluating text detoxification is motivated by the goal of fostering safer, more respectful online communication, rather than restricting freedom of expression. By focusing on developing better metrics for assessing detoxification quality, we aim to support the responsible deployment of detoxification models—ensuring that such systems are evaluated not only for effectiveness, but also for fairness, transparency, and contextual nuance.

We believe that detoxification tools, when applied, should serve as suggestions rather than enforcement mechanisms. Ideally, these tools would be integrated with user-centric interfaces that allow individuals to make informed decisions about their language use, with the final choice remaining in the hands of the user.

9. Acknowledgments

The work of Alexander Panchenko was supported by the RSF № 25-71-30008 "Laboratory for reliable, adaptive, and trustworthy Artificial Intelligence".

10. Bibliographical References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. *CoRR*, abs/2406.18403.

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for counterspeech against hate: A survey and how-to guide. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3480–3499, Mexico City, Mexico. Association for Computational Linguistics.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Maosong Cao, Alexander Lam, Haodong Duan, Hongwei Liu, Songyang Zhang, and Kai Chen. 2024. Compassjudger-1: All-in-one judge model helps model evaluation and evolution. *CoRR*, abs/2410.16256.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Society Open Science*, 5.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daryna Dementieva, Varvara Logacheva, Irina Nikishina, Alena Fenogenova, David Dale, I. Krotova, Nikita Semenov, Tatiana Shavrina, and Alexander Panchenko. 2022. RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora. In *Computational Linguistics and Intellectual Technologies (Dialogue-2022)*, Moscow, Russia.

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.

Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023. Exploring methods for cross-lingual text style transfer: The case of text detoxification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1083–1101, Nusa Dua, Bali. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Skyler Hallinan, Alisa Liu, Yejin Choi, and Maarten Sap. 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)*, pages 228–242, Toronto, Canada. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun. 2024. A multi-aspect framework for counter narrative evaluation using large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 147–168, Mexico City, Mexico. Association for Computational Linguistics.

Daniil Larionov, Mikhail Seleznyov, Vasiliy Viskov, Alexander Panchenko, and Steffen Eger. 2024. xCOMET-lite: Bridging the gap between efficiency and quality in learned MT evaluation metrics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21934–21949, Miami, Florida, USA. Association for Computational Linguistics.

Quentin Lemesle, Jonathan Chevelu, Philippe Martin, Damien Lolive, Arnaud Delhay, and Nelly Barbot. 2025. Paraphrase generation evaluation powered by an LLM: A semantic metric, not a lexical one. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8057–8087, Abu Dhabi, UAE. Association for Computational Linguistics.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *CoRR*, abs/2412.05579.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the 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 Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Zhongyu Luo, Man Luo, and Aiguo Wang. 2024. Multilingual text detoxification using google cloud translation and post-processing. 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 2769–2774. CEUR-WS.org.

Sourabrata Mukherjee, Akanksha Bansal, Atul Kr. Ojha, John P. McCrae, and Ondrej Dusek. 2023. Text detoxification as style transfer in English and Hindi. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 133–144, Goa University, Goa, India. NLP Association of India (NLPAI).

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 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)*, pages 189–194, Melbourne, Australia. Association for Computational Linguistics.

Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. 2025. A meta-evaluation of style and attribute transfer metrics. *CoRR*, abs/2502.15022.

Jiangao Peng, Zhongyuan Han, Huan Zhang, Jingyan Ye, Chang Liu, Biao Liu, Mingcan Guo, Haoyang Chen, Zijie Lin, and Yujiao Tang. 2024. A multilingual text detoxification method based on few-shot learning and CO-STAR framework. 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 2829–2837. CEUR-WS.org.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *WMT@EMNLP*.

Vitaly Protasov. 2024. PAN 2024 multilingual textdetox: Exploring cross-lingual transfer using large language models. 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 2852–2857. CEUR-WS.org.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C. Farinha, T. Glushkova, Alon Lavie, Luísa Coheur, and André F. T. Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Conference on Machine Translation*.

Ricardo Rei, Craig Alan Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *ArXiv*, abs/2009.09025.

Elisei Rykov, Konstantin Zaytsev, Ivan Anisimov, and Alexandr Voronin. 2024. Smurfcat at PAN 2024 textdetox: Alignment of multilingual transformers for text detoxification. 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 2866–2871. CEUR-WS.org.

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3178–3190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

A. Licensing of Resources

Below is an overview of the licenses associated with each resource used in this work (Table 1).

Resource	License	Homepage
TextDetoxEval-Style	OpenRail++	https://hf.co/datasets/textdetox/detoxification_pairwise_style_evaluation
TextDetoxEval-Content	OpenRail++	https://hf.co/datasets/textdetox/detoxification_pairwise_content_evaluation
TextDetoxEval-RU	OpenRail++	https://hf.co/datasets/textdetox/humaneval_textdetox_ru
Llama-pairwise-toxicity-evaluator	OpenRail++	https://hf.co/textdetox/Llama-pairwise-toxicity-evaluator
Llama-pairwise-content-evaluator	OpenRail++	https://hf.co/textdetox/Llama-pairwise-content-evaluator
LLaMa3	llama3	https://hf.co/meta-llama
DeepSeek	MIT	https://hf.co/collections/deepseek-ai/deepseek-r1-678e1e131c0169c0bc89728d
LaBSE	Apache 2.0	https://hf.co/sentence-transformers/LaBSE
xlmr-large-toxicity-classifier	Open RAIL++	https://hf.co/textdetox/xlmr-large-toxicity-classifier
XCOMET-lite	-	https://hf.co/myyycroft/XCOMET-lite
wmt22-comet-da	Apache 2.0	https://hf.co/Unbabel/wmt22-comet-da
XCOMET-XL	CC-BY-NC-SA-4.0	https://hf.co/Unbabel/XCOMET-XL
XCOMET-XXL	CC-BY-NC-SA-4.0	https://hf.co/Unbabel/XCOMET-XXL

Table 1: Overview of the licenses associated with each resource.

The licenses associated with the models and datasets used in this study are consistent with the intended purpose of conducting academic research aimed at advancing various NLP applications for positive impact.

B. LLMs Prompts

Here, we provide exact prompts used for LLMs prompting.

Fluency prompt

You are given with the text and your task is to evaluate the naturalness or fluency of the text.
Options:
yes (meaning that the text is natural and fluent)
partially (meaning that the text is partially natural and fluent)
no (meaning that the text is not natural and fluent at all)
Examples:
text: "I am going to the store to buy some groceries."
Answer: yes
text: "The cat sat on the mat dog argh adfsdf"
Answer: partially
text: "teh dsfsdf ats 1324@3 adf"
Answer: no
text: Answer:

Content similarity prompt

You are given with two texts. Your task is to compare the meaning of them and decide whether they mean the same or not.

Options:
- yes (both texts have the same meaning)
- no (there is significant difference in meaning between the two texts)

Examples:
text1: "Why do you behave like an asshole?"
text2: "Why do you act so bad?"
Answer: yes

text1: "This political pary is a gang of idiots"
text2: "They just dont know their stuff"
Answer: no

text1:
text2:
Answer:

`Style transfer evaluation prompt`

You are given with two texts. Your task is to compare which text is more toxic.
Options:
- text1 (text1 is more toxic than text2)
- text2 (text2 is more toxic than text1)
- None (both texts have equal toxicity - both of them are either toxic or non-toxic at the same time)
Examples:
text1: "I hate you, you are the worst person ever."
text2: "I love you, you are the best person ever."
Answer: text1
text1: "You are so smart, I can't believe you did that."
text2: "You are so stupid, I can't believe you did that."
Answer: text2
text1: "I think you are a great person."
text2: "I think you are a very good man"
Answer: none
You must return ONLY one of the options. Do not add any explanations or additional information.
text1: text2: Answer:

C. DialogueEvaluation-2022 Dataset

C.1. Dataset description

The DialogueEvaluation-2022 detoxification dataset was released as part of the first competition on Russian-language text detoxification. The dataset comprises toxic sentences sourced from Russian social media platforms such as Odnoklassniki, Pikabu, and Twitter. Data collection followed a three-stage crowdsourcing pipeline using Yandex.Toloka. First, annotators were asked to rewrite toxic sentences into fluent, non-toxic paraphrases while preserving the original meaning. Second, independent annotators verified that the paraphrases retained the same semantic content as the originals. Third, additional annotators checked whether the paraphrases were indeed non-toxic. Only paraphrases that passed both the content and toxicity checks (with $\geq$ 90% agreement) were included in the dataset.

Annotations were performed according to strict guidelines that defined toxicity and provided examples. Human evaluation focused on three components: *fluency* (grammatical correctness and readability), *content preservation* (semantic similarity with the original), and *style transfer quality* (removal of toxicity).

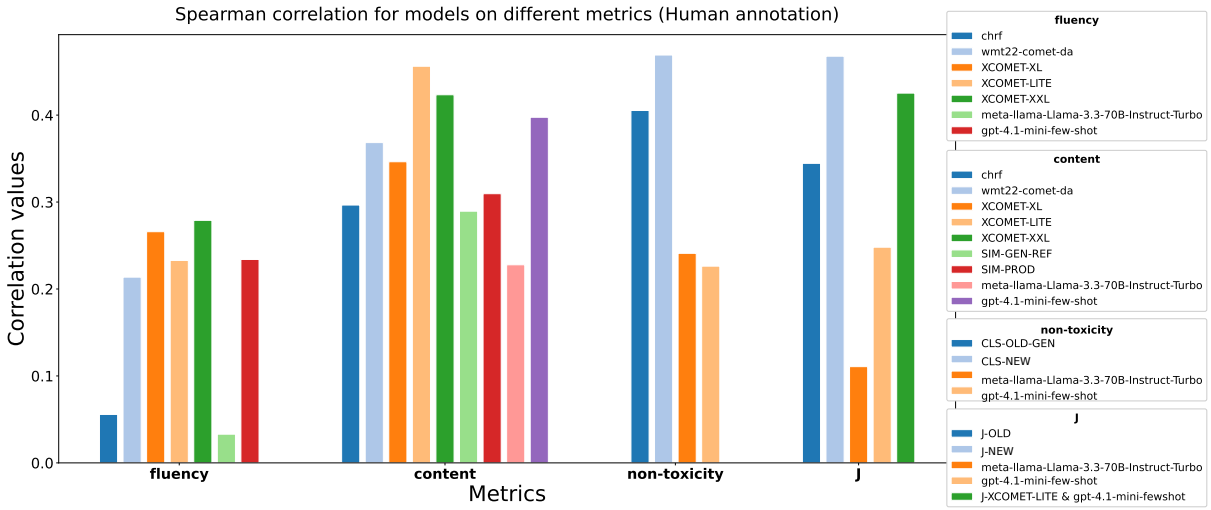

Figure 8: DialogueEvaluation-2022 results across different models.

C.2. Results

As we can see, for the **fluency** estimation, **XCOMET-XXL** still shows the best results, while **XCOMET-XL** performs slightly worse. At the same time, **gpt-4.1-mini-few-shot** demonstrates comparable results to **XCOMET-LITE**, though both fall short of the two aforementioned models.

Regarding **content similarity**, **XCOMET-LITE** achieves the best performance, while **gpt-4.1-mini-few-shot** yields results comparable to **XCOMET-XXL**.

For distinguishing between toxic and non-toxic texts, our new **CLS-PROD** model attains the highest scores, surpassing both the previous **CLS-OLD-GEN** model and the two considered LLMs. This is a surprising observation, as it directly contradicts the results from Figure 7 for the TEXTDETOXEVAL dataset, where our models performed significantly worse than nearly all LLMs for Russian. Finally, the joint evaluation scores are highest for the proposed model (**J-PROD**), outperforming all other models.