Overview of the Multilingual Text Detoxification Task at PAN 2025

Daryna Dementieva[1,*], Vitaly Protasov[2], Nikolay Babakov[3], Naquee Rizwan[4], Ilseyar Alimova[6], Caroline Brun[5], Vasily Konovalov[2], Arianna Muti[7], Chaya Liebeskind[8], Marina Litvak[9], Debora Nozza[7], Shehryaar Shah Khan[4], Sotaro Takeshita[10], Natalia Vanetik[9], Abinew Ali Ayele[11], Florian Schneider[12], Xintong Wang[12], Seid Muhie Yimam[12], Ashraf Elnagar[13], Animesh Mukherjee[4] and Alexander Panchenko[6,2]

[1]*Technical University of Munich, Munich, Germany*
[2]*Artificial Intelligence Research Institute, Moscow, Russia*
[3]*University of Santiago de Compostela, Santiago de Compostela, Spain*
[4]*Indian Institute of Technology, Kharagpur, India*
[5]*NAVER Labs Europe, Grenoble, France*
[6]*Skoltech, Moscow, Russia*
[7]*Bocconi University, Milan, Italy*
[8]*Jerusalem College of Technology, Jerusalem, Israel*
[9]*Shamoon Academic College of Engineering, Beer Sheva, Israel*
[10]*University of Mannheim, Mannheim, Germany*
[11]*Bahir Dar University, Bahir Dar, Ethiopia*
[12]*University of Hamburg, Hamburg, Germany*
[13]*University of Sharjah, Sharjah, UAE*

Abstract

Despite different countries and social platform regulations, digital abusive speech persists as a significant challenge. One of the way to tackle abusive, or more specifically, toxic language can be automatic text detoxification—a text style transfer task (TST) of changing register of text from toxic to more non-toxic. We extend our previous Multilingual Text Detoxification (TextDetox) task to new languages—Italian, French, Hebrew, Hinglish, Japanese, and Tatar—suggesting participants to participate in *multi-lingual* and *cross-lingual* text detoxification challenges. We provide insights into new data collection, evaluation metrics, as well as dive into the participants results.
Warning: This paper contains rude texts that only serve as illustrative examples.

Keywords

PAN 2025, Multilingual Text Detoxification, Text Style Transfer, Multilingualism

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
daryna.dementieva@tum.de (D. Dementieva); vitasprotas@gmail.com (V. Protasov); nikolay.babakov@usc.es (N. Babakov); nrizwan@kgpian.iitkgp.ac.in (N. Rizwan); alimovailseyar@gmail.com (I. Alimova); caroline.brun@naverlabs.com (C. Brun); vasily.konovalov@phystech.edu (V. Konovalov); arianna.muti@unibocconi.it (A. Muti); liebchaya@gmail.com (C. Liebeskind); marinal@sce.ac.il (M. Litvak); debora.nozza@unibocconi.it (D. Nozza); shehryaarshahkhan@gmail.com (S. S. Khan); sotaro.takeshita@uni-mannheim.de (S. Takeshita); natalyav@sce.ac.il (N. Vanetik); a.panchenko@skol.tech (A. Panchenko)
https://dardem.github.io (D. Dementieva); https://github.com/Vitaly-Protasov (V. Protasov); https://github.com/bbkjunior/bbkjunior (N. Babakov); https://www.linkedin.com/in/naquee-rizwan-a97abb159 (N. Rizwan); https://dblp.uni-trier.de/pid/211/4634.html (I. Alimova); https://europe.naverlabs.com/people_user_naverlabs/caroline-brun (C. Brun); https://github.com/vaskonov (V. Konovalov); https://faculty.unibocconi.eu/ariannamuti (A. Muti); https://www.jct.ac.il/facultyresearch/2017-2021/67/ (C. Liebeskind); https://en.sce.ac.il/faculty/marina-litvak (M. Litvak); https://www.deboranozza.com (D. Nozza); https://www.linkedin.com/in/shehryaar-shah-khan-4a18b9308 (S. S. Khan); https://sotaro.io/about (S. Takeshita); https://en.sce.ac.il/faculty/natalia-vanetik (N. Vanetik); https://scholar.google.com/citations?user=g2m1wH4AAAAJ&hl=en (A. A. Ayele); https://www.linkedin.com/in/flo-schneider-hh (F. Schneider); https://ethanscuter.github.io (X. Wang); https://seyyaw.github.io (S. M. Yimam); https://www.sharjah.ac.ae/en/academics/Colleges/CI/dept/cs/Pages/ppl_detail.aspx?mcid=4 (A. Elnagar); https://cse.iitkgp.ac.in/~animeshm (A. Mukherjee); https://alexanderpanchenko.github.io (A. Panchenko)

1. Introduction

While progress has been made in addressing **digital violence** [1] and abusive speech [2], there remains a pressing need for more **proactive approaches** to moderating hate and toxic language. In our TextDetox shared task, we continue to investigate **text detoxification** as a proactive method for toxic speech moderation, now extending our coverage to 15 languages: English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, Amharic, Italian, French, Hebrew, Hinglish, Japanese, and Tatar.

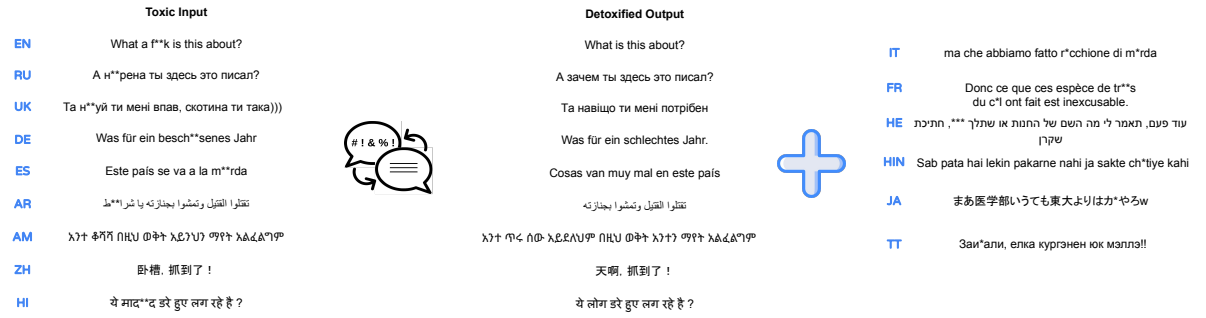

Figure 1: We suggested participants two challenges: multilingual text detoxification for 9 languages with parallel training data as well as cross-lingual transfer to new 6 languages.

In this shared task, we explored both setups—multilingual and cross-lingual one (Figure 1)—extending parallel text detoxification data from TextDetox 2024 to 6 new languages [3]. The remainder of this paper is structured as follows. Section 2 gives an overview of the TextDetox 2025 shared task rules. Section 3 provides the full overview of the new multilingual parallel text detoxification dataset collection per each language. In the following sections, the evaluation setups essentials are described—baselines in Section 4, automatic evaluation setup in Section 5, and LLM-as-a-judge evaluation setup in Section 6. The submissions from participants are described in Section 7. Section 8 provides the details about final results—both automatic (Section 8.1) and LLM-as-a-judge (Section 8.2) evaluation leaderboards. Finally, Section 9 concludes the paper.

All the resources produced from the task are listed at the shared task page[1] and are also mentioned in the corresponding sections. All the data, classifiers, and text detoxification baselines are released for a public usage at our HuggingFace space.[2] Also, we provide additional information on the annotation and detailed results at our Github repo for the corresponsing year.[3]

2. Shared Task Rules

The share task timeline was divided in to two phases—development and test.

Development Phase This year, together with already existing parallel data for English and Russian from previous works [4, 5], we released 400 parallel samples per each TextDetox 2024 language [6] as training data. Then, we used previous year languages 600 toxic samples as a test set as well as 100 toxic sentences per new 6 languages.

Test Phase We extended the test data for new languages to full 600 toxic samples as well. Thus, now, for all 15 languages, equal amount of toxic test instances are available. Participants were invited to submit *multilingual* and *cross-lingual* solutions.

[1] https://pan.webis.de/clef25/pan25-web/text-detoxification.html
[2] https://hf.co/textdetox
[3] https://github.com/textdetox/textdetox_clef_2025

Leaderboards During both phases, the leaderboards based on automatic evaluation were available. We used Codalab platform [7].[4] At each phase leaderboard, we highlighted scores per each challenge—**AvgP** for the languages with parallel training data available and **AvgNP** for new languages without any training data. This year, we also significantly improved the **automatic evaluation** metrics (Section 5). At the same time, for additional leaderboard, we provided as well **LLM-as-a-judge** results with fine-tuned LLMs for text detoxification evaluation task (Section 6). Participants were asked to analyze their performance from both leaderboards.

3. Multilingual Parallel Text Detoxification Dataset

Firstly, we re-used the data from TextDetox 2024 shared task [6] for 9 languages—English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, Amharic—as: (i) 400 parallel sentences per each language now were used as a training data for all phases; (ii) 600 toxic sentences per each language served as a part of the test sets for both dev and test phases.

Then, for new languages—Italian, French, Hebrew, Hinglish, Japanese, and Tatar—we asked experts and native speakers to contribute for new corpora collection. Further, we describe the collection details per each language: , French (Section 3.1), Italian (Section 3.2), Hebrew (Section 3.3), Japanese (Section 3.4), Hinglish (Section 3.5), and Tatar (Section 3.6).

For all the data collection, we adapt the concept of English ParaDetox [4] collection pipeline. The quality check consists of three main criteria:

Task 1: Rewrite text in a polite way Annotators need to provide the detoxified paraphrase of the text so it becomes non-toxic and the main content is saved or to skip paraphrasing if the text is not possible to rewrite in non-toxic way;

Task 2: Do these sentences mean the same? Check if the content is indeed the same between the original toxic text and its potential non-toxic paraphrase;

Task 3: Is this text offensive? Verification of the provided paraphrase if it is indeed non-toxic.

In the same manner, each language stakeholder asked the annotators to rewrite the toxic samples verifying the main three criteria: (i) the new paraphrase should be non-toxic; (ii) the content should be saved as much as possible; (iii) the resulted text should be fluent but may contain some minor mistakes (as the majority of the original toxic samples are examples from posts from social networks).

We explicitly communicated to language stakeholders that deletion of toxic words should be considered only as a last resort in the detoxification process—used solely when rephrasing is not feasible. Annotators were instructed to prioritize **rephrasing toxic segments** wherever possible, relying on deletion only when no suitable neutral alternative could be constructed.

For new languages, we obtained 600 parallel pairs from which toxic parts were revealed as dev (first 100) and test (full 600) sets.

3.1. French

We introduce the DetoxifyFR dataset, a novel detoxification dataset for French, comprising 600 toxic comments and their human-written neutral rewrites, incorporated into the test phase of the shared task.

3.1.1. Input Data Preparation

The DetoxifyFR dataset is constructed from two distinct sources:

[4] hhttps://codalab.lisn.upsaclay.fr/competitions/22396

FrenchToxicityPrompts [8]: 50,000 naturally occurring French samples, annotated with toxicity scores from the *Perspective API*. This data originates from Lélu, a French dialogue dataset extracted from Reddit's public French datasets. Conversations are segmented into sentences using *spaCy*, with *Perspective API* scores ranging from 0 (not toxic) to 100 (highly toxic) assigned to each sentence. We retain toxic and highly toxic sentences (i.e., scores ≥ 50) as candidates for detoxification, resulting in 12,601 sentences.

Jigsaw Multilingual Toxic Comment Classification test set [9]: 9,274 French samples annotated as toxic or non-toxic, from which we retain 1,557 toxic samples as candidates for detoxification.

3.1.2. Annotation Process

LLM-based Pre-filtering Producing a neutral version of a toxic sentence while preserving its meaning is not always feasible, as some sentences are inherently toxic and can not be detoxified without a drastic change in content. To streamline the manual annotation process, we prompted Llama-3.1-70B-Instruct to evaluate whether a sample could be detoxified. Only sentences deemed suitable for detoxification were passed to the next phase. This LLM-based filtering retained 1,062 toxic candidates from Jigsaw and 9,103 from FrenchToxicityPrompts. Interestingly, on FrenchToxicityPrompts, we observe that the LLM filters out approximately 50% of the highly toxic sentences (toxicity ≥ 75) but only 22% of the toxic sentences ($50 \leq$ toxicity < 75).

Manual Annotation We then randomly selected comments, evenly split between the two filtered datasets. The annotation process is entirely manual and does not rely on LLM-generated content. Initial tests with LLMs for generating detoxified sentences revealed biases in the annotations, leading us to adopt a fully manual approach. A total of 600 French samples, approximately evenly distributed from the two data sources, were detoxified during this process.

Annotator The annotator is a native French speaker with extensive experience in linguistic data annotation and holds a PhD in computational linguistics.

LLM-based Validation We applied a validation step on the final data using an LLM-based evaluation. For this, we prompted Qwen2.5-72B-Instruct to assess the toxicity score of the 600 detoxified sentences on a 5-point scale (from 0: not offensive to 4: toxic) and content preservation (from 0: same content to 4: different content). Results are presented in Tables 1 and 2.

Table 1
Percentage of neutral sentences per toxicity score assessed by LLM judge.

Score	Description	Percentage of sentences
0	Not offensive	96%
1	Mildly offensive	3.7%
2	Moderately offensive	0.3%
3	NA	0%
4	NA	0%

The LLM judge demonstrates high-quality detoxified data, with 96% of sentences rated non-offensive (Table 1). Additionally, Table 2 shows strong content preservation, with 61.6% of sentences having only slight tone changes and 24.3% exhibiting reduced emotional intensity (e.g., lowered aggressiveness), indicating effective detoxification while largely maintaining content integrity.

Table 2
Percentage of neutral sentences per content preservation score assessed by LLM judge.

Score	Description	Percentage of sentences
0	Same content	10.8%
1	Slight change in tone	61.6%
2	Reduced emotional intensity	24.3%
3	Content not fully preserved	3%
4	Content not preserved	0.3%

3.2. Italian

We introduce the DetoxifyIT dataset, a new detoxificaton dataset for Italian, featuring 600 toxic comments and their human-written neutral rewrites.

3.2.1. Input Data Preparation

We use three datasets for manual detoxification: two from Twitter and one from Wikipedia. The Twitter datasets originate from EVALITA shared tasks—AMI (2020), focused on misogyny [10], and HODI (2023), targeting homotransphobia [11], each comprising approximately 5,000 annotated tweets. Posts are labeled as either hate or non-hate speech, with additional subcategories provided for hateful content. The Wikipedia dataset is drawn from Jigsaw's Multilingual Toxic Comment Classification Challenge[5], and consists entirely of toxic comments.

To identify content suitable for manual detoxification, we applied a multi-stage filtering process. All datasets were constrained to a post length of 5–30 words to ensure contextual clarity. Since hate speech and toxicity labels do not fully overlap — for instance, non-toxic content may still be hateful and vice versa — we additionally used the Perspective API[6] to obtain toxicity scores for the Twitter data. We excluded tweets that were either insufficiently toxic or excessively severe, ultimately retaining 400 tweets, with 200 per target group. A detailed description of the filtering methodology is provided in [12].

3.2.2. Annotation Process

The annotation process followed the guidelines established by the 2024 edition of this shared task [6], with the primary objectives of removing toxicity while preserving the original meaning. Annotators were instructed to rephrase toxic content wherever possible, using deletion only as a last resort. Three native Italian speakers with expertise in NLP and toxic language worked on the rewrites. The process was collaborative: one annotator rewrote the first set of 100 texts, and then all three reviewed them together to ensure consistency with the guidelines. This group review was repeated after 300 and 600 texts. The final dataset includes only texts on which all three annotators agreed. A fourth expert later reviewed the full set and suggested small improvements where needed.

3.3. Hebrew

We introduce the **HeDetox** dataset, a new detoxification dataset for Hebrew constructed from offensive online forum comments and annotated through a multi-stage process. Our approach builds on prior linguistic taxonomies and recent prompting techniques for detoxification.

[5]https://kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification
[6]https://www.perspectiveapi.com/

3.3.1. Input Data Preparation

The Hebrew HeDetox dataset is derived from user-generated content on a popular Israeli news forum (https://rotter.net/forum/listforum.php), where emotionally charged discussions on current events are frequent. A custom web scraping pipeline was used to collect discussion threads, extract metadata (e.g., timestamps, post IDs), and normalize the comment text. A strict anonymization process removed personally identifiable information such as usernames, mentions, and embedded links.

To identify toxic content, we employed a few-shot classification method using large language models (LLMs), prompted with definitions and reasoning chains based on the *Simplified Offensive Language (SOL) Taxonomy* [13]. This taxonomy provides a stepwise classification structure that includes offense type, target, vulgarity level, and implicit linguistic devices (e.g., irony, metaphor). Each comment was annotated by the LLM as explicitly offensive, implicitly offensive, or non-offensive. To ensure precision, we retained only those labeled as explicitly offensive, discarding borderline or ambiguous examples.

Input Toxicity Data The toxic inputs for HeDetox were selected from the scraped corpus based on LLM classifications and filtered for explicit offensiveness to serve as input for the detoxification task.

3.3.2. Annotation Process

Annotation Tasks We adapted the few-shot Chain-of-Thought (CoT) prompting framework introduced by Dementieva et al. [14] for Hebrew. A custom Hebrew-language prompt was designed to guide the LLM in identifying elements of toxicity and producing neutralized rewrites. The prompt included keyword-based reasoning, strict instructions to preserve meaning and tone, and negative examples that illustrated undesirable behaviors such as unsolicited advice or paraphrasing. We applied the prompt in a few-shot setting, providing two in-context examples before detoxifying each offensive sentence.

Manual Correction of LLM Outputs To assess the quality of LLM-generated detoxifications, we implemented a two-phase manual correction process. In the first phase, 100 sentences were independently reviewed and revised by two annotators, with a third adjudicator resolving discrepancies and ensuring adherence to annotation guidelines. Annotators were instructed to avoid common issues such as over-softening, omission of key content, introduction of new information, imprecise synonym use, and retention of toxic language.

Inter-annotator agreement was evaluated using cosine similarity over sentence embeddings. We compared heBERT [15], multilingual BERT (mBERT) [16], and traditional vector models (n-grams, tf-idf). Table 3 presents the results. Despite syntactic variability, both heBERT and mBERT demonstrated

Table 3
Original and detoxified sentences similarity.

Representation	Cosine Similarity
heBERT SE	0.888
mlBERT SE	0.937
n-grams	0.649
tf-idf	0.685

strong semantic agreement between annotators, while traditional syntactic representations showed lower similarity.

In the second phase (500 sentences), we streamlined the process by assigning one annotator and one corrector per sentence. The annotator generated detoxified rewrites based on the same guidelines as in phase one, and the corrector reviewed them to ensure semantic fidelity and minimal stylistic alteration. This setup reduced variation and improved consistency across the dataset.

Annotators Our annotators are native speakers that have previous hate speech annotation experience and hold degrees in Computer Science and Software Engineering. One of the annotators is male, and one is female.

3.4. Japanese

The Japanese split is constructed by a CS graduate student who is a native Japanese speaker in the following procedure:

3.4.1. Input Data Preparation

We base our data construction on the open2ch corpus [17],[7] a large collection of user-generated texts from a popular thread-based social platform in Japan covering various topics. We apply keyword-based filtering to obtain sentences that are likely to be toxic as our starting point for the annotation.[8] This filtering is applied to 10,000 sentences from the original dataset, and approximately 60% of the sentences are detected as potentially toxic and compose the dataset for annotation described in the subsequent section.

3.4.2. Annotation Process

Given the dataset of potentially toxic sentences, a native Japanese speaker from a CS PhD program carried out the annotation. The semantics from the original texts are preserved as much as possible, and the samples are omitted when (I) The whole text is toxic, making it unable to detoxify, (II) The original text is not toxic at all. At the end of the annotation, 3488 sentences are considered, and 600 sentences have been detoxified; other sentences could not be annotated due to the two previously mentioned reasons. Most of the unannotated samples are invalid because of the reason (I). This is due to the nature of the data source: a long-running, fully anonymized thread-based online platform.

3.5. Hinglish

3.5.1. Input Data Preparation

Input Toxicity Data We used the aggression annotated corpus of Hindi-English code-mixed data proposed in [18] for framing 600 samples of toxic-detoxified pairs. Contents in the dataset are obtained from *Facebook* and are made up of a combination of Hindi-English code-mixed posts that are relevant within Indian subcontinent. Publicly available dataset[9] consists of two splits– *train* and *dev*. We sampled our data from *train*; *dev* split and remaining samples from *test* were used to train toxicity classifier. Publicly available *train* split contains a total of 12,000 posts bifurcated into three categories OAG *(overtly aggressive)*, CAG *(covertly aggressive)* and NAG *(non-aggressive)* each containing 4856, 4869 and 2275 samples respectively. Since our work is centered on text detoxification, we carefully sample 600 posts from OAG and CAG categories that have toxic contents and are detoxifiable.

Input Preprocessing From an initial collection of 9,725 posts across the OAG and CAG categories, we first filtered out posts written in Hindi, leaving us exclusively with samples in Hinglish. We then performed deduplication through exact string matching to eliminate duplicate entries. To ensure data cleanliness and consistency, mentions, links, and emojis were systematically removed. Posts containing fewer than five tokens, separated by whitespace, were excluded from the dataset, while those exceeding twenty-five tokens were reformulated to adhere to this length constraint. Importantly, all these modifications were made in a way that preserved the original intent and toxicity levels of each post.

[7] https://hf.co/datasets/p1atdev/open2ch
[8] https://github.com/MosasoM/inappropriate-words-ja
[9] https://github.com/victor7246/Hinglish_Hate_Detection/blob/main/data/raw/trac1-dataset/hindi/agr_hi_train.csv

3.5.2. Annotation Process

Annotation Task(s) After the initial input preprocessing, posts were meticulously reviewed through manual verification and were systematically categorized into two groups: detoxifiable and non-detoxifiable. This classification, along with the rephrasing process discussed in previous section, was conducted by an NLP researcher with practical expertise in hate speech and toxic language mitigation.

From the preprocessed dataset of 4,824 posts, a total of 600 detoxifiable posts were collected. Collection was halted once this target was reached. From these, a representative subset of 20 posts was carefully selected as benchmark detoxification samples. These samples underwent expert review by two native Hindi speakers to ensure high-quality reference standards. Using these expert-validated examples as guidance, annotators were instructed to rephrase toxic content into non-toxic expressions while preserving the original meaning of each post. The detoxification process was carried out independently by two trained annotators, whose details are provided in a dedicated subsection. Each sample was detoxified by one annotator.

Annotators A male NLP researcher, with expertise in the detection and mitigation of hate speech and toxic language, was engaged alongside a third-year undergraduate male student with practical experience in machine learning. Both individuals are native Hindi speakers from India and possess an in-depth understanding of the thematic content encompassed within the dataset. Furthermore, their strong proficiency in reading and writing Hinglish ensures precise and nuanced annotation. Together, they are responsible for executing a comprehensive detoxification of the entire dataset, leveraging their specialized skills and linguistic expertise.

3.6. Tatar

We introduce the **TatDetox** dataset, a high-quality, fully manually annotated and validated detoxification dataset for Tatar constructed from social media posts.

3.6.1. Input Data Preparation

The data was sourced from the Web Corpus, a collection that gathers texts from various resources focusing on the minority languages of Russia.[10] For the Tatar language, the Web Corpus provides posts on the social network VKontakte. Two methods were applied for data filtering and selecting examples for detoxification. The initial filtering was based on a toxic lexicon [6]. We used both Tatar and Russian lexicons, as the texts feature code-switching and many obscene words used in Tatar are borrowed from Russian. Additionally, a sentiment classifier was used [19]. The classifier identifies six emotions: Anger, Joy, Sadness, Fear, Disgust, and Surprise. It was observed that the largest number of toxic texts appeared in the Anger class. Thus, statements labeled with this category were selected for annotation.Texts obtained through different filtering methods were combined, cleaned of HTML tags, anonymized, and then sent for further annotation.

3.6.2. Annotation Process

Two annotators participated in the annotation process, both native speakers, one of whom is an expert in the field of NLP. The annotators were tasked with checking the text for toxicity and writing a detoxified version for toxic examples following the general guidelines provided by task organizers. The annotators were also instructed to preserve the original spelling: if the text was written exclusively in Russian letters, the rewritten version had to use only Russian letters as well; if Tatar letters were used, the detoxified text needed to maintain this writing system. Cross-validation of the examples was then carried out, with each annotator validating the examples provided by the other. A total of 1004 examples were selected for annotation, of which 600 were included in the final dataset.

[10] http://web-corpora.net/wsgi3/minorlangs/download

4. Baselines

We provide five baselines for our shared task: (i) a trivial Duplicate baseline, (ii) a rule-based Delete approach, (iii) a Backtranslation pipeline that reduces the task to a monolingual setting, (iv) a fine-tuned mT0 model covering 9 out of the 15 languages, and (v) zero-shot LLMs with instruction prompts. The code for all baselines is publicly available[11].

Duplicate A trivial baseline where the output is simply a copy of the input. No detoxification is applied, and the original toxic content is returned unchanged.

Delete This unsupervised baseline removes toxic or obscene substrings from the input text based on predefined keyword lists. For the shared task, we compiled such lists for all 15 target languages using publicly available resources (see Table 4). The number of keywords varies by language, reflecting morphological diversity and differing ways of expressing toxicity. We release the full multilingual keyword collection for participants and public use[12].

Table 4
The list of the original sources and the corresponding amount of obscene keywords used to compile multilingual toxic lexicon list for the Delete baseline.

Language	Original Source	# of Keywords
Amharic	Ours+[20]	245
Arabic	Ours+[20]	430
German	[21, 20]	247
English	[4, 22, 20]	3 390
Spanish	[20]	1 200
Hindi	[20]	133
Russian	[5, 20]	141 000
Ukrainian	[23, 20]	7 360
Chinese	[24, 25, 20]	3 840
Italian	[26, 21]	815
Japanese	[27]	328
Hebrew	[28]	731
French	[21, 29, 30]	1 290
Tatar	[26]+translated Russian	15 600
Hinglish	[31]	209

Backtranslation This is a more sophisticated unsupervised baseline based on cross-lingual transfer. The approach works by first translating non-English inputs into English using the NLLB-3.3B model [20].[13] Detoxification is then performed using the English-language model bart-base-detox, fine-tuned on the ParaDetox training set [4].[14] Finally, the detoxified text is translated back into the original target language using NLLB. For Hinglish, we use the specialized RLM-hinglish-translator model.[15]

Fine-tuned mT0 We consider the mT0-XL-Detox-ORPO model [32],[16] one of the top-performing systems from the TextDetox 2024 shared task. Given its strong performance, we adopt it as a baseline

[11] https://github.com/pan-webis-de/pan-code/tree/master/clef25/text-detoxification/baselines
[12] https://hf.co/datasets/textdetox/multilingual_toxic_lexicon
[13] https://hf.co/facebook/nllb-200-3.3B
[14] https://hf.co/s-nlp/bart-base-detox
[15] https://hf.co/rudrashah/RLM-hinglish-translator
[16] https://hf.co/s-nlp/mt0-xl-detox-orpo

in this year's competition. Although it was fine-tuned on only 9 of the 15 languages considered in this competition, we observed promising zero-shot performance on the 6 remaining languages.

LLMs Prompting This baseline uses the Llama-3.1-70B-Instruct model,[17] with few-shot examples embedded in the instruction prompt. The model performs detoxification based on the given examples without any task-specific fine-tuning. This baseline also relies on general-purpose instruction prompts but uses various proprietary OpenAI models, including GPT-4 (0613), GPT-4o (2024-08-06), and GPT-3.5 (o3-mini-2025-01-31). We provide a basic detoxification prompt and evaluate the zero-shot capabilities of these models without additional tuning (we utilized same prompt[18] for all models).

5. Automatic Evaluation Setup

We adopt the evaluation approach from CLEF 2024 competition [6] and apply adjustments to the underlying models. As in that work, we also measure the final detoxification metric as a combination of three sub-metrics: toxicity, content similarity and fluency of the generated text compared to source and reference texts. The evaluation script is available online.[19]

Toxicity Measurement (TOX) assesses how toxic the evaluated text is. Since the goal is to detoxify the text, lower scores indicate better performance. We evaluate toxicity using an XLM-R [33] model fine-tuned on a multilingual toxicity corpus covering 15 languages. We released both the model[20] and the underlying dataset[21]. The model was fine-tuned using supervised learning (SFT) for a binary classification task on around 5 000 samples per each language sampled from many corpora used for text detoxification data selection. For calculating final toxicity score, we specifically consider the model's predicted probability that the generated text belongs to the «toxic» class. In contrast to the evaluation made in the CLEF-2024 competition, we not only updated the model but also adjusted the approach for calculating the final fluency score. This was done by comparing the probability that the generated text, source input text, and reference output text belong to the «toxic» class, using the following rules: (i) if the probability of the generated text is higher than for the source input, we penalize the score and set it to **zero**; (ii) if the probability of the generated text is lower than for the reference text, we reward the score and set it to **one**.

Content similarity (SIM) evaluates how well the generated texts preserve key semantic information from the original input. This metric penalizes outputs that miss essential content during the detoxification process. Following CLEF-2024, we compute content similarity using cosine similarity between LaBSE[22] embeddings [34]. However, unlike the previous setup, which only measures similarity between the source input and the generated text—ignoring reference outputs—our approach addresses this limitation. We propose an improved metric that combines both input-output and output-reference similarities, using a weighted sum. This enhancement captures not only fidelity to the original input but also alignment with human-annotated reference texts, providing a more comprehensive evaluation of content preservation: $\textbf{SIM}(s_i, g_i, r_i) = \textbf{cos}(s_i, g_i) * \mathbf{w}_{s_i,g_i} + \textbf{cos}(g_i, r_i) * \mathbf{w}_{g_i,r_i}$, where $\mathbf{w}_{s_i,g_i} + \mathbf{w}_{g_i,r_i} = 1$.

Fluency Estimation (FL) measures how natural, coherent, and grammatically correct a generated text is—essentially, how closely it resembles language produced by a native speaker. In the context of generated detoxification, this reflects whether the output reads smoothly and idiomatically without spelling mistakes or unnatural constructions. In CLEF-2024, fluency was measured using

[17] https://hf.co/mlabonne/Llama-3.1-70B-Instruct-lorablated
[18] https://github.com/pan-webis-de/pan-code/tree/master/clef25/text-detoxification/baselines/openai
[19] https://github.com/pan-webis-de/pan-code/tree/master/clef25/text-detoxification
[20] https://hf.co/textdetox/xlmr-large-toxicity-classifier-v2
[21] https://hf.co/datasets/textdetox/multilingual_toxicity_dataset
[22] https://hf.co/sentence-transformers/LaBSE

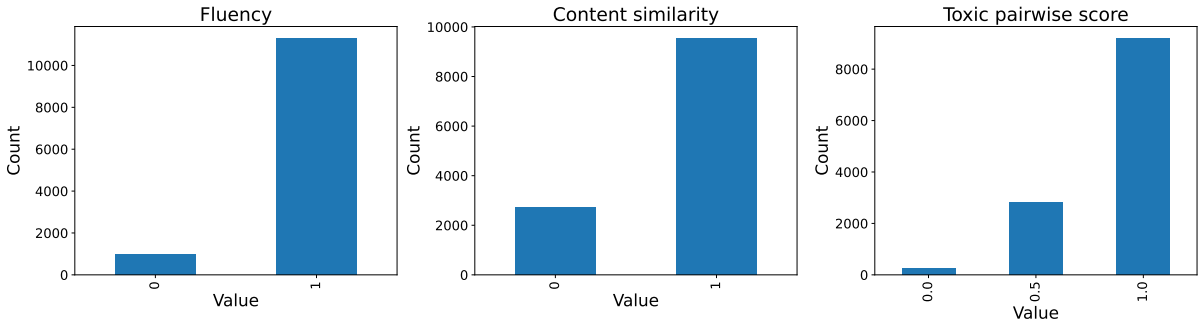

Figure 2: Distribution of training labels available for fine-tuning the LLM. The scores come from human evaluations of pairs consisting of a toxic sentence and its detoxified paraphrase. Fluency: 0 = less fluent than original, 1 = equally fluent. Content similarity: 0 = different content, 1 = same content. Toxicity pairwise score (scoreing generated paraphrase): 0 = more toxic than original, 0.5 = equally toxic, 1 = less toxic

ChrF [35] scores between generated outputs and human-annotated references. However, this approach ignores the original toxic input, leading to a bias toward reference-style outputs and neglecting the relevance of the transformation from source to target. To address this, we adopt XCOMET [36], a metric originally designed for machine translation evaluation. Unlike ChrF, XCOMET considers the input–generation–reference triplet, modeling fluency in the context of both the source and reference. It leverages pretrained language models to assess fluency beyond surface-level matching, incorporating deeper semantic and syntactic patterns. We specifically use *myyycroft/XCOMET-lite* [37], a compressed version of *Unbabel/XCOMET-XXL* [36] that retains over 95% of its performance while reducing computational cost by 60%. This efficiency makes it suitable for real-time evaluation in our competition platform and for participant use.

Joint score (J) is the aggregation of the three aforementioned metrics.

$$\mathbf{J} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{TOX}(s_i, g_i, r_i) \cdot \mathbf{SIM}(s_i, g_i, r_i) \cdot \mathbf{FL}(s_i, g_i, r_i),$$

where $\mathbf{TOX}(s_i, g_i, r_i)$, $\mathbf{SIM}(s_i, g_i, r_i)$, $\mathbf{FL}(s_i, g_i, r_i) \in [0, 1]$ for each text detoxification output g_i, source toxic text s_i and reference annotated detoxification text r_i.

6. LLM as a Judge

As an additional evaluation strategy, we employed LLMs as automatic judges to assess the quality of system submissions. We explored two main paradigms: the use of out-of-the-box, pre-trained LLMs, and customized LLMs fine-tuned specifically for the evaluation tasks. The out-of-the-box models utilized in our experiments included GPT-4.1 mini, GPT-4.1 nano, CompassJudger-1-32B-Instruct [38], DeepSeek-R1-Distill-Qwen-32B [39], DeepSeek-V3-0324 [39], and Llama-3.3-70B-Instruct.[23] The prompts used for all models for three tasks are shown in Appenidx B.

For model customization and further alignment with the requirements of the shared task, we conducted additional fine-tuning experiments using Llama-3.1-8B[24] and Qwen-3-8B [40][25] models. Fine-tuning was performed using the Low-Rank Adaptation (LoRA) [41] method to efficiently adapt the base models while minimizing computational overhead. The main model weights were loaded in 4-bit quantized format, enabling faster training and reduced memory usage without significant loss in performance. The LoRA configuration utilized the following hyperparameters: rank $r = 8$, $\alpha = 16$, and a dropout rate of 0.1, with adaptation applied to all linear layers. Optimization was carried out using

[23] https://hf.co/meta-llama/Llama-3.3-70B-Instruct
[24] https://hf.co/meta-llama/Llama-3.1-8B
[25] https://hf.co/Qwen/Qwen3-8B

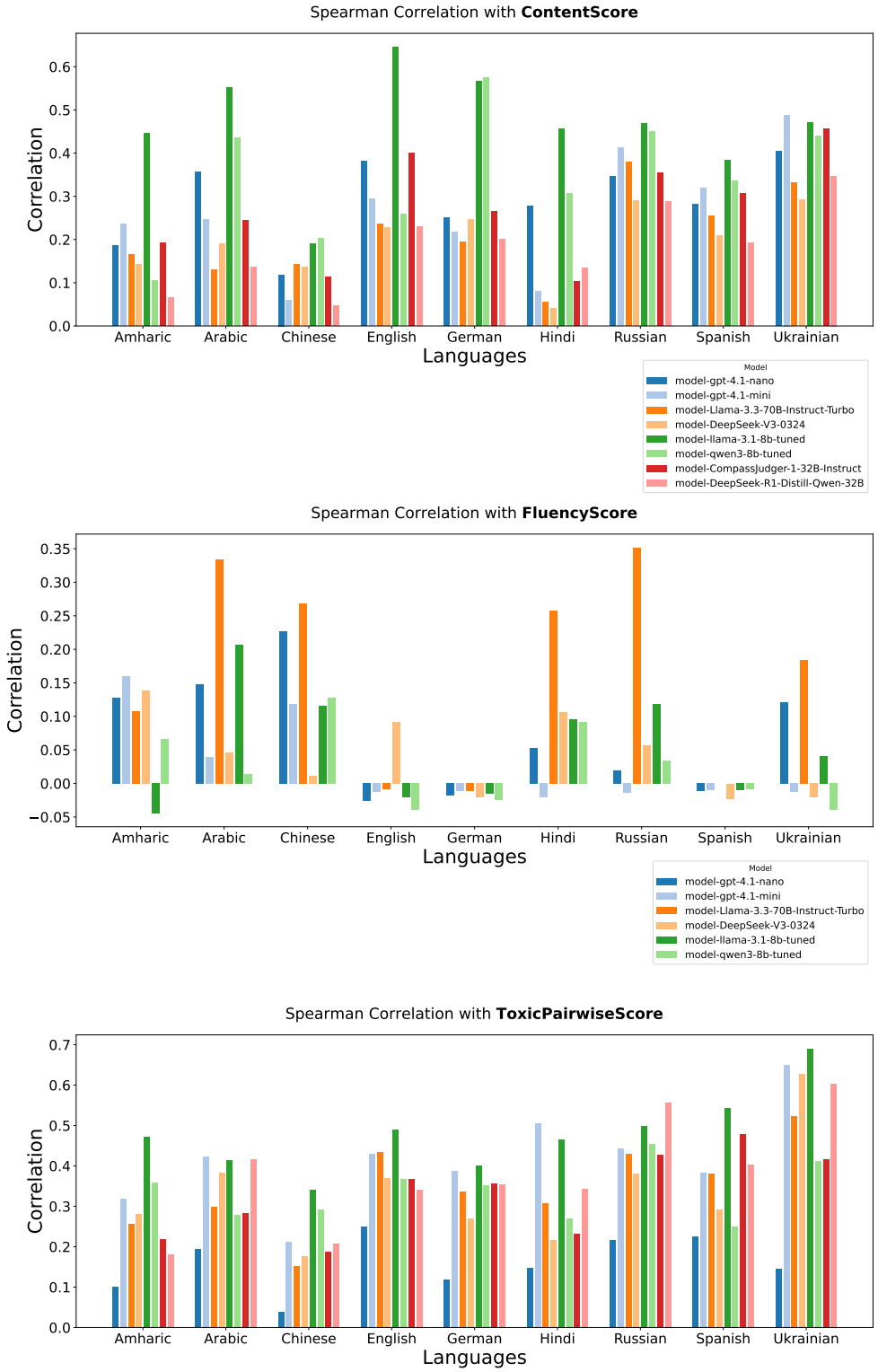

Figure 3: Correlation of different LLMs with human judgments on the tasks of toxic pairwise and content similarity scoring.

a learning rate of 1×10^{-4}, cosine learning rate scheduler, weight decay of 1×10^{-4}, warmup ratio set to 0.0, and maximum gradient norm of 1.0. 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 evaluation models were trained (where applicable) and tested using datasets from the CLEF

TextDetox 2024 [6] and RUSSE 2022 [42, 43] competitions, both of which provide pairs of toxic and detoxified paraphrases annotated with human quality scores. For our experiments, we utilized a subset of the CLEF TextDetox 2024 dataset as the test set, while the remaining data from both CLEF 2024 and RUSSE 2022 were used for training and development. This resulted in a total of 12,279 training pairs and 4,320 test pairs.

Figure 2 presents the distribution of the labels available for fine-tuning within the selected dataset. It is evident that the data is skewed towards positive cases. While this imbalance is somewhat justifiable for the toxic pairwise score and content similarity score, in the case of the fluency score, the skew is particularly severe: the dataset contains 984 negative and $11,295$ positive instances. We conducted an initial round of fine-tuning experiments for both of the selected models using the fluency score. However, the extreme class imbalance resulted in models that consistently predicted only the positive class, rendering the approach ineffective. Consequently, we decided not to pursue further fine-tuning experiments for fluency and do not report results for this setting. Therefore, our fine-tuning efforts focused exclusively on content similarity and toxic pairwise scoring, while fluency is reported only for the out-of-the-box LLMs.

The results of our experiments are presented in Figure 3. It is evident that, for most languages, the fine-tuned Llama-3.1-8B outperforms the alternative LLMs on the tasks of content similarity and toxic pairwise scoring. Based on these results, we selected the fine-tuned Llama 3.1-8B as the primary model for these tasks.

With respect to fluency, all models evaluated in their out-of-the-box configurations demonstrated consistently low correlation with human judgments (below 0.35). As a result, we do not consider any of the tested models suitable for the fluency assessment task in our shared task setting.

7. Participants

We received 25 submissions for the development phase leaderboard and 26 submissions for the test phase leaderboard. Here, we briefly describe some solutions of our final participants:

Sky.Duan [44] Employs a parallel architecture that integrates both local models and large language models for multilingual text detoxification. The system combines s-nlp/mt0-xl-detox-orpo and Qwen3 to leverage the strengths of specialized and general-purpose models, enabling robust and intelligent detoxification across different languages.

d1n910 [45] Utilized a Chain-of-Thought (CoT) prompting approach with the Deepseek-r1 large language model to enhance the reasoning capabilities and effectiveness of text detoxification.

Pratham [46] Developed a multilingual text detoxification system centered on MT0-XL with task-specific prompting, complemented by language-specific lexical filtering using custom toxic word dictionaries. This hybrid approach ensured robust detoxification across 15 languages, particularly handling challenges in code-mixed and morphologically rich languages through handcrafted filtering rules.

ducanhhbtt [47] Built an efficient multilingual detoxification system leveraging the Gemma 3 12B model with LoRA-based fine-tuning and advanced prompting, including few-shot retrieval and chain-of-thought reasoning. The approach combined progressive fine-tuning phases and extensive data augmentation to ensure high performance across both high- and low-resource languages, all while maintaining computational efficiency.

nikita.sushko [48] Applied supervised fine-tuning over the MultiParaDetox, SynthDetoxM, and a custom synthetic dataset generated using the SynthDetoxM pipeline to enhance detoxification performance.

humairafaridq [49] Proposed a prompt-driven, truly multilingual approach using only the GPT-4o-mini model and in-context learning, where each toxic input is paired with a fixed instruction and three language-specific toxic-to-neutral examples, eliminating the need for model fine-tuning.

Jiaozipi [50] Introduced a multilingual detoxification method based on an ensemble of large language models (DeepSeek, Qwen, Kimi) guided by the RISE framework and hint engineering, using few-shot examples and multi-dimensional evaluation to select optimal outputs without fine-tuning.

Oleg_Papulov Utilized Qwen3-0.6B with LoRA fine-tuning on neural-toxic text pairs to achieve detoxification.

SVATS [51] Explored multiple model architectures (Qwen2-7B, Gemma-2 4B), comparing full and LoRA fine-tuning, dataset variations, and multilingual vs. English-only strategies, while also evaluating few-shot prompting with GPT-4o and a baseline deletion method for robust multilingual detoxification.

MetaDetox [52] Eliminated the need for fine-tuning by applying Chain-of-Thought prompting and few-shot learning with DeepSeek, generating stylistically diverse rewrites for each input, and selecting the best output via semantic similarity and toxicity-based reranking across 15 languages.

The Toxinators 2000 (jellyproll) Used the baseline MT0 model for most languages, while applying a vocabulary replacement method for Hinglish and a combination of vocab replacement and MT0 for Tatar and Japanese.

Team Detox (Gopal) [53] Developed a hybrid system combining rule-based toxic span masking with few-shot prompting of GPT-4o-mini, where toxic words are masked and both masked and original sentences are provided to the model to generate detoxified outputs.

Nililusu (ylmmcl) [54] Created a multilingual pipeline integrating lexicon- and classifier-based toxicity detection, translation for non-English inputs, and ensemble detoxification using three generative models, with outputs selected by evaluation metrics and back-translated to the original language.

wl2776 Trained T5 model using the prompt "Detoxify: <sentence>" to perform text detoxification.

SomethingAwful Utilized Llama 3.1 with explicit reasoning for generation, combined with a best-of-five selection strategy using example-based Self-BLEU Scoring (SBS) to choose the optimal detoxified output.

8. Results

Here, we provide the final results of the final test phase of our shared task for both automatic leaderboard from Codalab (Section 8.1) and LLM-as-a-Judge (Section 8.2). The full detailed tables of results per each language and per each metric set can be found in Appendix A.

8.1. Automatic Evaluation Leaderboard

The results of the automatic evaluation used in Codalab for languages with parallel training data (AvgP) are presented in Table 5, for new languages without parallel training data—in Table 6.

Several diverse teams succeeded in surpassing multiple baseline models presented in our shared task. However, unlike in the previous edition, no team was able to outperform the human reference outputs this year. Among the submissions, Team **ducanhhbtt** with Gemma model underneath achieved the highest overall scores for languages with available training data, closely followed by Team **MetaDetox**,

sky.Duan, and **Team Pratham**. Notably, these were the only teams that consistently outperformed our strong mT0 baseline, which had been fine-tuned on the provided multilingual training data. Interestingly, the mT0 baseline also significantly outperformed the GPT-4 prompting-based baseline, highlighting that effective text detoxification still requires either model fine-tuning or advanced prompting strategies—prompting alone appears insufficient.

Despite strong overall performance, no single team demonstrated consistent success across all languages. For instance, *Team Pratham* led specifically in Hindi and Russian, *adugeen* excelled in Ukrainian, and *Jiaozipi* delivered top performance in Spanish and Hindi. These outcomes reinforce the ongoing challenge of building multilingual models that generalize robustly across all languages, even when parallel training data is available.

Table 5
Results of the *automatic* evaluation of the test phase for 9 languages **with parallel data**. Scores are sorted by the average **J**oint score. The scores for each language are respective **J**oint scores. Baselines are highlighted with gray, Human References are highlighted with green. Top-3 best scores for each language are highlighted with **bold**, the best score is **underlined bold**.

Team	AvgP*	AM	AR	DE	EN	ES	HI	RU	UK	ZH
Human References	0.854	0.742	0.869	0.936	0.822	0.863	0.841	0.867	0.899	0.850
ducanhhbtt	**0.6852**	0.446	0.718	**0.798**	0.734	0.686	0.619	0.749	0.799	**0.618**
Team MetaDetox	0.6850	0.415	**0.732**	0.766	**0.742**	**0.719**	0.629	0.753	0.798	0.611
sky.Duan	0.6764	0.491	0.715	0.757	0.727	0.696	0.627	0.754	0.770	0.551
Team Pratham	0.6759	0.486	0.724	0.750	0.729	0.696	**0.6314**	**0.755**	0.776	0.533
baseline_mt0	0.675	0.491	0.715	0.757	0.727	0.696	0.627	0.754	0.770	0.543
jellyproll	0.675	0.491	0.715	0.757	0.727	0.696	0.627	0.754	0.770	0.543
adugeen	0.670	0.440	0.713	0.769	0.685	0.709	0.619	0.750	**0.800**	0.543
Jiaozipi	0.656	0.369	0.682	0.748	0.724	0.712	0.6313	0.730	0.773	0.539
SVATS	0.656	0.461	0.668	0.754	0.704	0.698	0.593	0.725	0.766	0.531
baseline_gpt4	0.637	0.412	0.603	0.728	0.708	0.708	0.605	0.706	0.747	0.513
humairafaridq	0.636	0.373	0.621	0.729	0.726	0.688	0.584	0.703	0.754	0.546
nikita.sushko	0.628	0.437	0.612	0.727	0.716	0.666	0.594	0.657	0.738	0.509
ylmmcl	0.612	0.448	0.643	0.684	0.667	0.581	0.578	0.690	0.693	0.523
Gopal	0.611	0.364	0.648	0.707	0.583	0.662	0.573	0.665	0.725	0.570
d1n910	0.604	0.387	0.589	0.693	0.632	0.673	0.583	0.681	0.733	0.461
baseline_o3mini	0.562	0.291	0.498	0.607	0.688	0.660	0.549	0.638	0.685	0.439
SomethingAwful	0.549	0.205	0.508	0.618	0.647	0.652	0.580	0.627	0.689	0.418
baseline_delete	0.536	0.461	0.611	0.586	0.473	0.603	0.480	0.514	0.581	0.516
baseline_backtranslation	0.481	0.265	0.438	0.513	0.684	0.528	0.419	0.696	0.498	0.290
baseline_duplicate	0.475	0.461	0.564	0.572	0.353	0.566	0.417	0.424	0.442	0.477

For the newly introduced languages, the performance landscape shifted notably—often quite dramatically—compared to the original set. Starting with the baselines, the mT0 model, which was fine-tuned only on the original nine languages, experienced a substantial drop in performance and was clearly outperformed by GPT-4. This outcome is expected: mT0 lacks exposure to the new languages, whereas GPT-4, as a more general-purpose large language model, has likely seen a broader range of languages

during pretraining. The only exception was Italian, where mT0 performed comparably well—likely due to the presence of Spanish (a closely related language) in the training data.

Team **ducanhhbtt** once again led the leaderboard, now achieving the highest average score (AvgNP) across five of the six new languages, with only Tatar posing a challenge. For Tatar, **jellyproll** outperformed other teams by incorporating targeted vocabulary substitution specifically tailored to this underrepresented language. Notably, Team **adugeen** showed a significant performance boost on the new languages, rising to second place overall.

An important takeaway is that many more teams were able to surpass the best-performing GPT-4 baseline this year. This demonstrates that general-purpose LLMs, while powerful, are not sufficient out-of-the-box for specialized tasks like multilingual text detoxification—particularly in low-resource settings. The top-performing teams effectively integrated insights from our shared task into their pipelines, leveraging techniques such as cross-lingual transfer, vocabulary adaptation, and advanced prompting. These results underscore the need for task- and language-specific adaptation to ensure robust and culturally sensitive content moderation across diverse linguistic contexts.

Table 6
Results of the *automatic* evaluation of the test phase for 6 languages **without parallel data**. Scores are sorted by the average **J**oint score. The scores for each language are respective **J**oint scores. Baselines are highlighted with gray, Human References are highlighted with green. Top-3 best scores for each language are highlighted with **bold**, the best score is **underlined bold**.

Team	AvgNP*	IT	JA	HE	FR	TT	HIN
Human References	0.847	0.918	0.884	0.772	0.905	0.825	0.781
ducanhhbtt	**0.643**	**0.784**	**0.674**	**0.531**	**0.8020**	0.556	**0.511**
adugeen	0.623	0.761	0.640	0.479	0.785	0.582	0.491
Team MetaDetox	0.609	0.755	0.587	0.530	0.8019	0.498	0.481
Jiaozipi	0.607	0.728	0.647	0.499	0.801	0.485	0.480
jellyproll	0.605	0.746	0.644	0.415	0.760	**0.617**	0.449
SVATS	0.599	0.755	0.589	0.451	0.769	0.573	0.455
Gopal	0.595	0.746	0.662	0.478	0.700	0.516	0.470
baseline_gpt4	0.579	0.742	0.637	0.513	0.780	0.468	0.333
humairafaridq	0.578	0.700	0.663	0.489	0.740	0.452	0.422
d1n910	0.5753	0.698	0.637	0.507	0.744	0.450	0.416
Team Pratham	0.5747	0.749	0.591	0.416	0.752	0.584	0.356
baseline_mt0	0.572	0.746	0.582	0.415	0.760	0.580	0.351
nikita.sushko	0.512	0.727	0.495	0.449	0.754	0.386	0.262
SomethingAwful	0.511	0.637	0.496	0.473	0.717	0.352	0.392
baseline_delete	0.510	0.668	0.441	0.436	0.518	0.573	0.425
sky.Duan	0.501	0.663	0.573	0.446	0.699	0.427	0.196
baseline_o3mini	0.484	0.605	0.490	0.475	0.725	0.360	0.251
baseline_duplicate	0.482	0.653	0.440	0.425	0.447	0.510	0.419
ylmmcl	0.471	0.645	0.528	0.323	0.625	0.492	0.213
baseline_backtranslation	0.342	0.462	0.241	0.339	0.626	0.254	0.133

8.2. LLM as a Judge Leaderboard

The results of the LLM-as-a-Judge evaluation for languages with parallel training data (AvgP) are presented in Table 7, for new languages without parallel training data—in Table 8.

With the updated evaluation setup, the overall ranking of systems shifted slightly, and new leading teams emerged for several languages. Interestingly, as detailed in the full evaluation reports in the Appendix A, LLM-as-a-Judge results show that, for languages such as English, Spanish, Russian, Ukrainian, and Hebrew, some newly submitted systems significantly outperformed even the human references. This trend reflects both the relative resource richness of certain languages—where modern models likely benefit from more extensive exposure to toxicity-related data—and the emergence of underrepresented languages where LLMs, due to their broader pretraining, now demonstrate greater fluency and generalization capabilities compared to older decoder-based models.

Table 7
Results from the *post* evaluation of the test phase for 9 languages **with parallel data**. Scores are sorted by the average **J**oint score. The scores for each language are respective **J**oint scores. Baselines are highlighted with gray, Human References are highlighted with green. Top-3 best scores for each language are highlighted with **bold**, the best score is **underlined bold**.

Team	AvgP*	AM	AR	DE	EN	ES	HI	RU	UK	ZH
Human References	0.822	0.742	0.838	0.930	0.846	0.783	0.807	0.783	0.780	0.887
Team MetaDetox	**0.812**	0.626	**0.826**	**0.9191**	**0.893**	**0.823**	**0.785**	0.829	0.7906	**0.813**
ducanhhbtt	0.798	0.614	0.814	0.9189	0.871	0.797	0.762	0.827	0.785	0.796
adugeen	0.775	0.597	0.79	0.888	0.794	0.765	0.773	0.792	**0.7911**	0.783
Jiaozipi	0.768	0.579	0.779	0.84	0.882	0.812	0.701	0.817	0.757	0.744
Team Pratham	0.768	0.621	0.793	0.822	0.843	0.763	0.758	0.811	0.782	0.717
baseline_mt0	0.768	0.639	0.791	0.825	0.843	0.764	0.751	0.809	0.77	0.717
jellyproll	0.768	**0.6384**	0.788	0.828	0.842	0.758	0.759	0.818	0.766	0.715
humairafaridq	0.768	0.526	0.768	0.912	0.888	0.814	0.731	0.805	0.747	0.724
sky.Duan	0.765	0.6382	0.787	0.83	0.847	0.757	0.747	0.811	0.771	0.696
d1n910	0.742	0.591	0.72	0.866	0.805	0.807	0.689	0.792	0.757	0.652
nikita.sushko	0.735	0.491	0.652	0.828	0.858	0.74	0.732	**0.835**	0.772	0.702
SVATS	0.723	0.38	0.705	0.854	0.83	0.749	0.672	0.798	0.743	0.776
Gopal	0.722	0.58	0.718	0.819	0.691	0.757	0.701	0.792	0.742	0.699
baseline_gpt4	0.715	0.482	0.686	0.807	0.858	0.8	0.647	0.778	0.723	0.654
ylmmcl	0.714	0.594	0.718	0.772	0.796	0.624	0.716	0.76	0.727	0.725
baseline_o3mini	0.676	0.421	0.595	0.747	0.893	0.796	0.609	0.711	0.663	0.652
SomethingAwful	0.663	0.367	0.592	0.749	0.856	0.763	0.631	0.685	0.694	0.629
baseline_delete	0.558	0.499	0.61	0.564	0.453	0.543	0.566	0.583	0.577	0.63
baseline_backtranslation	0.458	0.425	0.442	0.479	0.743	0.466	0.395	0.689	0.256	0.231
baseline_duplicate	0.432	0.38	0.446	0.479	0.37	0.451	0.432	0.45	0.455	0.429

In the final AvgP results, team **MetaDetox** secured first place, demonstrating top performance across all nine languages with training data. They are closely followed by team **ducanhhbtt**, while team **adugeen** now ranks third, maintaining the strongest performance for Ukrainian. Additionally, team **jellyproll** achieved the highest score for Amharic, and team **nikita.sushko** led in Russian.

For the new languages without training data, team **adugeen** now holds the highest average score, with **ducanhhbtt** following closely behind. Impressively, **ducanhhbtt** achieved near top results across all six new languages, aligning with earlier findings from the automatic evaluation. A notable development is the strong performance by team **humairafaridq**, which unexpectedly placed third overall for new languages. Significant shifts were observed particularly in the rankings for Hebrew and Hinglish, highlighting the dynamic nature of model generalization in low-resource and culturally specific settings.

We congratulate all participating teams for their creative and impactful solutions—several of which surpassed even proprietary systems like GPT-4. While this year's evaluation setup showed stronger alignment with human judgments, the results also underscore the continued need for developing more robust, culturally aware evaluation metrics for multilingual text style transfer.

Table 8
Results from the *post* evaluation of the test phase for 6 languages **without parallel data**. Scores are sorted by the average **J**oint score. The scores for each language are respective **J**oint scores. Baselines are highlighted with gray, Human References are highlighted with green. Top-3 best scores for each language are highlighted with **bold**, the best score is **underlined bold**.

Team	AvgNP*	IT	JA	HE	FR	TT	HIN
Human References	0.783	0.893	0.904	0.528	0.888	0.724	0.716
adugeen	**0.722**	0.823	0.805	0.657	0.86	0.583	0.606
ducanhhbtt	0.72	**0.842**	**0.82**	**0.681**	**0.889**	0.495	0.592
humairafaridq	0.718	0.819	0.819	0.671	0.883	0.511	0.586
Gopal	0.704	0.812	0.784	0.631	0.843	0.575	0.578
Team MetaDetox	0.691	0.821	0.721	0.61	0.883	0.493	**0.621**
d1n910	0.674	0.791	0.796	0.581	0.853	0.436	0.584
Jiaozipi	0.688	0.795	0.787	0.611	0.850	0.541	0.544
sky.Duan	0.668	0.822	0.769	0.619	0.873	0.416	0.509
baseline_gpt4	0.662	0.79	0.779	0.578	0.865	0.438	0.524
SVATS	0.662	0.775	0.734	0.576	0.815	0.523	0.549
jellyproll	0.648	0.742	0.745	0.495	0.79	**0.611**	0.503
baseline_mt0	0.641	0.749	0.711	0.501	0.793	0.598	0.494
Team Pratham	0.639	0.752	0.71	0.495	0.801	0.575	0.502
nikita.sushko	0.628	0.763	0.722	0.573	0.807	0.492	0.41
baseline_o3mini	0.559	0.748	0.661	0.497	0.826	0.209	0.411
SomethingAwful	0.579	0.728	0.643	0.49	0.814	0.324	0.477
ylmmcl	0.53	0.662	0.647	0.357	0.659	0.543	0.312
baseline_delete	0.525	0.628	0.443	0.496	0.576	0.521	0.486
baseline_duplicate	0.429	0.455	0.442	0.407	0.46	0.421	0.387
baseline_backtranslation	0.254	0.333	0.147	0.349	0.503	0.054	0.139

9. Conclusion

In Multilingual Text Detoxification shared task at PAN 2025, participants were tasked with transforming text style from toxic to non-toxic across 15 languages: English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian, Amharic, Italian, French, Hebrew, Hinglish, Japanese, and Tatar. The task was divided into two challenges: *multilingual* one with the parallel training data available as well as *cross-lingual* one stressing models with new unseen languages. Participants' submissions in both phases underwent evaluation using an improved set of automatic metrics, followed by additional novel LLM-as-a-Judge evaluation.

We received a wide range of submissions leveraging both LLMs—such as DeepSeek, Gemma, Qwen, and GPT-4—as well as fine-tuned decoder models like mT0, often enhanced with specialized preprocessing or combined methodologies. Many participating teams succeeded in outperforming even our strongest baselines, including fine-tuned on our data mT0 and GPT-4. These results highlight the importance of adapting standard LLM prompting or fine-tuning approaches to the unique demands of the text detoxification task, especially when dealing with underrepresented languages. Notably, several systems even surpassed human reference outputs in languages such as English, Spanish, Russian, Hebrew, and Ukrainian.

This year, we also introduced a more comprehensive evaluation framework that combined improved automatic metrics with an additional LLM-as-a-Judge setup. By comparing both leaderboards, we observed consistent overall trends with slight shifts in ranking, though certain teams and languages exhibited significant reordering. These variations reflect the inherent complexity and subjectivity of tasks like text detoxification and proactive moderation of abusive speech—challenges that are deeply influenced by linguistic and cultural context. The results emphasize the continued need for high-quality human-annotated data and more sophisticated, culturally sensitive automatic evaluation metrics to ensure fair and reliable evaluation process.

Acknowledgment

We express our deepest gratitude to Toloka.ai platform for our shared task support. Daryna Dementieva's work was additionally supported by Alexander Fraser's TUM Heilbronn chair as well as Friedrich Schiedel TUM Think Tank Fellowship. Naquee Rizwan, Shehryaar Shah Khan, and Animesh Mukherjee would like to thank SPARC-II (Scheme for Promotion of Academic and Research Collaboration, Phase II) project for funding international travel and subsistence to carry out this work. Ilseyar Alimova would like to thank AIRI for funding the preparation of the Tatar dataset and Dina Abdullina for help with dataset annotation. Arianna Muti's and Debora Nozza's research is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101116095, PERSONAE). Arianna Muti and Debora Nozza are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

Declaration on Generative AI

During this study, AI assistant was utilized in the writing process. ChatGPT was employed for paraphrasing throughout the paper's formulation followed by authors thorough additional verification.

References

[1] Z. R. Shi, C. Wang, F. Fang, Artificial intelligence for social good: A survey, CoRR abs/2001.01818 (2020). URL: http://arxiv.org/abs/2001.01818. arXiv:2001.01818.

[2] N. Rizwan, S. M. Yimam, D. Dementieva, F. Skupin, T. Fischer, D. Moskovskiy, A. A. Borkar, R. Geislinger, P. Saha, S. Roy, M. Semmann, A. Panchenko, C. Biemann, A. Mukherjee, HATEPRISM:

Policies, platforms, and research integration. advancing nlp for hate speech proactive mitigation, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, Vienna, Austria, 2025. URL: https://www.inf.uni-hamburg.de/en/inst/ab/lt/publications/2025-rizwanetal-acl-hateprism.pdf.

[3] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[4] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, A. Panchenko, ParaDetox: Detoxification with parallel data, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 6804–6818. URL: https://aclanthology.org/2022.acl-long.469. doi:10.18653/v1/2022.acl-long.469.

[5] D. Dementieva, V. Logacheva, I. Nikishina, A. Fenogenova, D. Dale, I. Krotova, N. Semenov, T. Shavrina, A. Panchenko, RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora, COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES (2022). URL: https://api.semanticscholar.org/CorpusID:253169495.

[6] D. Dementieva, D. Moskovskiy, N. Babakov, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. Ustalov, E. Stakovskii, A. Smirnova, A. Elnagar, A. Mukherjee, A. Panchenko, Overview of the multilingual text detoxification task at PAN 2024, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), 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*, CEUR-WS.org, 2024, pp. 2432–2461. URL: https://ceur-ws.org/Vol-3740/paper-223.pdf.

[7] A. Pavao, I. Guyon, A. Letournel, D. Tran, X. Baró, H. J. Escalante, S. Escalera, T. Thomas, Z. Xu, Codalab competitions: An open source platform to organize scientific challenges, J. Mach. Learn. Res. 24 (2023) 198:1–198:6. URL: https://jmlr.org/papers/v24/21-1436.html.

[8] C. Brun, V. Nikoulina, FrenchToxicityPrompts: a large benchmark for evaluating and mitigating toxicity in French texts, in: R. Kumar, A. K. Ojha, S. Malmasi, B. R. Chakravarthi, B. Lahiri, S. Singh, S. Ratan (Eds.), Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024, ELRA and ICCL, Torino, Italia, 2024, pp. 105–114. URL: https://aclanthology.org/2024.trac-1.12/.

[9] I. Kivlichan, J. Sorensen, J. Elliott, L. Vasserman, M. Görner, P. Culliton, Jigsaw multilingual toxic comment classification., https://kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification, 2020.

[10] E. Fersini, D. Nozza, P. Rosso, AMI @ EVALITA2020: automatic misogyny identification, in: V. Basile, D. Croce, M. D. Maro, L. C. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020, volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2765/paper161.pdf.

[11] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the first shared task on homotransphobia detection in italian, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3473/paper26.pdf.

[12] V. De Ruvo, A. Muti, D. Dementieva, D. Nozza, Detoxify-it: An italian parallel dataset for text detoxification, in: The 9th Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, 2025.

[13] B. Lewandowska-Tomaszczyk, A. Baczkowska, O. Dontcheva-Navrátilová, C. Liebeskind, G. V. Oleškevičienė, S. Žitnik, M. Trojszczak, R. Povolná, L. Selmistraitis, A. Utka, D. Gudelis, Llod schema for simplified offensive language taxonomy in multilingual detection and applications, Lodz Papers in Pragmatics 19 (2023) 301–324. URL: https://doi.org/10.1515/lpp-2023-0016. doi:doi:10.1515/lpp-2023-0016.

[14] D. Dementieva, N. Babakov, A. Ronen, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. Moskovskiy, E. Stakovskii, E. Kaufman, A. Elnagar, A. Mukherjee, A. Panchenko, Multilingual and explainable text detoxification with parallel corpora, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 7998–8025. URL: https://aclanthology.org/2025.coling-main.535/.

[15] A. Chriqui, I. Yahav, Hebert & hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition, INFORMS Journal on Data Science (2022).

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.

[17] M. Inaba, おーぷん2ちゃんねる対話コーパスを用いた用例ベース対話システム, in: 第87回言語・音声理解と対話処理研究会(第10回対話システムシンポジウム), 人工知能学会研究会資料SIG-SLUD-B902-33, 2019, pp. 129–132.

[18] R. Kumar, A. N. Reganti, A. Bhatia, T. Maheshwari, Aggression-annotated corpus of Hindi-English code-mixed data, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1226.

[19] S. H. Muhammad, N. Ousidhoum, I. Abdulmumin, J. P. Wahle, T. Ruas, M. Beloucif, C. de Kock, N. Surange, D. Teodorescu, I. S. Ahmad, D. I. Adelani, A. F. Aji, F. D. M. A. Ali, I. Alimova, V. Araujo, N. Babakov, N. Baes, A. Bucur, A. Bukula, G. Cao, R. T. Cardenas, R. Chevi, C. I. Chukwuneke, A. Ciobotaru, D. Dementieva, M. S. Gadanya, R. Geislinger, B. Gipp, O. Hourrane, O. Ignat, F. I. Lawan, R. Mabuya, R. Mahendra, V. Marivate, A. Piper, A. Panchenko, C. H. P. Ferreira, V. Protasov, S. Rutunda, M. Shrivastava, A. C. Udrea, L. D. A. Wanzare, S. Wu, F. V. Wunderlich, H. M. Zhafran, T. Zhang, Y. Zhou, S. M. Mohammad, BRIGHTER: bridging the gap in human-annotated textual emotion recognition datasets for 28 languages, CoRR abs/2502.11926 (2025). URL: https://doi.org/10.48550/arXiv.2502.11926. doi:10.48550/ARXIV.2502.11926. arXiv:2502.11926.

[20] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Y. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation, CoRR abs/2207.04672 (2022). URL: https://doi.org/10.48550/arXiv.2207.04672. doi:10.48550/ARXIV.2207.04672. arXiv:2207.04672.

[21] I. Shutterstock, List of dirty, naughty, obscene, and otherwise bad words, https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words, 2020. Accessed: 2023-12-12.

[22] R. J. Gabriel, English full list of bad words and top swear words banned by google, https://github.com/coffee-and-fun/google-profanity-words/blob/main/data/en.txt, 2023. Accessed: 2023-12-12.

[23] K. Bobrovnyk, The dictionary of ukrainian obscene words, https://github.com/saganoren/obscene-ukr, 2019. Accessed: 2023-12-12.

[24] A. Jiang, X. Yang, Y. Liu, A. Zubiaga, SWSR: A chinese dataset and lexicon for online sexism detection, Online Soc. Networks Media 27 (2022) 100182. URL: https://doi.org/10.1016/j.osnem.2021.100182. doi:10.1016/J.OSNEM.2021.100182.

[25] J. Lu, B. Xu, X. Zhang, C. Min, L. Yang, H. Lin, Facilitating fine-grained detection of Chinese

toxic language: Hierarchical taxonomy, resources, and benchmarks, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023, pp. 16235–16250. URL: https://aclanthology.org/2023.acl-long.898.

[26] Meta Research, Toxicity-200, 2023. URL: https://github.com/facebookresearch/flores/blob/main/toxicity/README.md, accessed July 2025.

[27] K. Hashimoto, 概要, https://github.com/MosasoM/inappropriate-words-ja, 2020.

[28] C. Liebeskind, M. Litvak, N. Vanetik, From linguistics to practice: a case study of offensive language taxonomy in Hebrew, in: Y.-L. Chung, Z. Talat, D. Nozza, F. M. Plaza-del Arco, P. Röttger, A. Mostafazadeh Davani, A. Calabrese (Eds.), Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 110–117. URL: https://aclanthology.org/2024.woah-1.8/. doi:10.18653/v1/2024.woah-1.8.

[29] Wiktionary contributors, Catégorie: Insultes en français, 2021. URL: https://fr.wiktionary.org/wiki/Cat%C3%A9gorie:Insultes_en_fran%C3%A7ais, accessed July 2025.

[30] Wiktionary contributors, Catégorie: Termes vulgaires en français, 2021. URL: https://fr.wiktionary.org/wiki/Cat%C3%A9gorie:Termes_vulgaires_en_fran%C3%A7ais, accessed July 2025.

[31] P. Mathur, R. Sawhney, M. Ayyar, R. R. Shah, Did you offend me? classification of offensive tweets in hinglish language, in: D. Fiser, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, J. Wernimont (Eds.), Proceedings of the 2nd Workshop on Abusive Language Online, ALW@EMNLP 2018, Brussels, Belgium, October 31, 2018, Association for Computational Linguistics, 2018, pp. 138–148. URL: https://doi.org/10.18653/v1/w18-5118. doi:10.18653/V1/W18-5118.

[32] E. Rykov, K. Zaytsev, I. Anisimov, A. Voronin, Smurfcat at PAN 2024 textdetox: Alignment of multilingual transformers for text detoxification, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), 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*, CEUR-WS.org, 2024, pp. 2866–2871. URL: https://ceur-ws.org/Vol-3740/paper-276.pdf.

[33] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://doi.org/10.18653/v1/2020.acl-main.747. doi:10.18653/V1/2020.ACL-MAIN.747.

[34] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 878–891. URL: https://doi.org/10.18653/v1/2022.acl-long.62. doi:10.18653/V1/2022.ACL-LONG.62.

[35] M. Popovic, chrf: character n-gram f-score for automatic MT evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal, The Association for Computer Linguistics, 2015, pp. 392–395. URL: https://doi.org/10.18653/v1/w15-3049. doi:10.18653/V1/W15-3049.

[36] N. M. Guerreiro, R. Rei, D. van Stigt, L. Coheur, P. Colombo, A. Martins, xcomet: Transparent machine translation evaluation through fine-grained error detection, Transactions of the Association for Computational Linguistics 12 (2023) 979–995. URL: https://api.semanticscholar.org/CorpusID:264146484.

[37] D. Larionov, M. Seleznyov, V. Viskov, A. Panchenko, S. Eger, xCOMET-lite: Bridging the gap between efficiency and quality in learned MT evaluation metrics, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 21934–21949. URL: https://aclanthology.org/2024.emnlp-main.1223.

[38] M. Cao, A. Lam, H. Duan, H. Liu, S. Zhang, K. Chen, Compassjudger-1: All-in-one judge model helps model evaluation and evolution, CoRR abs/2410.16256 (2024). URL: https://doi.org/10.48550/arXiv.2410.16256. doi:10.48550/ARXIV.2410.16256. arXiv:2410.16256.

[39] DeepSeek-AI, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL: https://arxiv.org/abs/2501.12948. arXiv:2501.12948.

[40] Q. Team, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[41] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, 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, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[42] D. Dementieva, V. Logacheva, I. Nikishina, A. Fenogenova, D. Dale, I. Krotova, N. Semenov, T. Shavrina, A. Panchenko, RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora, COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES (2022). URL: https://api.semanticscholar.org/CorpusID:253169495.

[43] D. Dementieva, N. Babakov, A. Panchenko, MultiParaDetox: Extending text detoxification with parallel data to new languages, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 124–140. URL: https://aclanthology.org/2024.naacl-short.12.

[44] D. Xianbing, H. Zhongyuan, P. Jiangao, S. Kaiyin, Multilingual Text Detoxification System Based on Parallel Architecture: An Intelligent Approach Integrating Local Models and Large Language Models, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[45] J. Peng, S. Kaiyin, L. Kaichuan, L. Zhankeng, H. Zhongyuan, A Multilingual Text Detoxification Method Based on Chain-of-Thoughts Prompting Approach, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[46] P. Shah, V. Shah, S. Kale, Multilingual Text Detoxification via Prompted MT0-XL and Lexical Filtering, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[47] T. D. A. Dang, F. P. D'Elia, GemDetox: Enhancing a massively multilingual model for text detoxification on low-resource languages, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[48] A. Voronin, D. Moskovsky, N. Sushko, PAN 2025 Textdetox: Exploring a Sage-T5-like approach for text detoxification, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[49] H. Farid, Z. Ahmad, A. Mahmood, I. Ameer, HF_Detox at PAN 2025 TextDetox: Prompt-Driven Multilingual Detoxification, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[50] X. Liu, Y. Yi, Z. Chen, S. Xu, Z. Ke, X. Guo, Y. Huang, W. Zhang, J. Chen, Y. Han, Jiaozipi at CLEF 2025:A Multilingual Text Detoxification Method Based on Large Language Model-Based Ensemble Learning, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[51] V. Kozlovskiy, A. Ploskin, S. Tantry, T. Matveeva, S. Savelyeva, Can Small Models Outperform Large Ones in Text Detoxification?, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[52] S. Bourbour, A. S. Kelishami, M. Gheysari, F. Rahimzadeh, Cross-Lingual Detoxification with Few-Chain Prompting: A Competitive System for TextDetox 2025, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[53] N. Krishna, L. Sai Teja, A. Mishra, Team Detox at PAN: Multilingual Text Detoxification using LLM, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[54] N. Lai-Lopez, S. Yuan, L. Wang, L. Zhang, Lexicon-Guided Detoxification and Classifier-Gated Rewriting: A PAN 2025 Submission, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

A. Automatic and LLM-as-a-Judge Full Evaluation Results

Here, we provide the extended results—from both automatic and LLM evaluation setups—based on three evaluation parameters for all languages: Amharic (Table 9), Arabic (Table 10), German (Table 11), English (Table 12), Spanish (Table 13), Hindi (Table 14), Russian (Table 15), Ukrainian (Table 16), Chinese (Table 17), French (Table 18), Hebrew (Table 19), Hinglish (Table 20), Italian (Table 21), Japanese (Table 22), and Tatar (Table 23). In every table, the baselines are highlighted with gray; Human References are highlighted with green; the ordering is made by **J** score from **LLM-as-a-Judge Evaluation** results. The automatic evaluation is based on the full test set of 600 samples per language; LLM evaluation was performed on 100 set of the test set per language.

Table 9
Automatic and LLM evaluation results for Amharic.

	Automatic Evaluation				LLM evaluation			
	FL	**SIM**	**STA**	**J**	**FL**	**SIM**	**STA**	**J***
Human References	0.846	0.873	1.0	0.742	0.846	0.888	0.981	0.742
baseline_mt0	0.737	0.772	0.836	0.491	0.737	0.997	0.878	0.639
sky.Duan	0.737	0.772	0.836	0.491	0.737	0.997	0.876	0.638
jellyproll	0.737	0.772	0.836	0.491	0.737	0.997	0.876	0.638
Team MetaDetox	0.678	0.727	0.814	0.415	0.678	0.997	0.928	0.626
Team Pratham	0.743	0.785	0.81	0.486	0.743	0.995	0.847	0.621
ducanhhbtt	0.677	0.714	0.895	0.446	0.677	0.962	0.943	0.614
adugeen	0.698	0.735	0.828	0.44	0.698	0.948	0.896	0.597
ylmmcl	0.698	0.724	0.851	0.448	0.698	0.95	0.89	0.594
d1n910	0.647	0.652	0.89	0.387	0.647	0.948	0.957	0.591
Gopal	0.65	0.732	0.735	0.364	0.65	0.992	0.909	0.58
Jiaozipi	0.627	0.661	0.859	0.369	0.627	0.952	0.96	0.579
humairafaridq	0.706	0.773	0.659	0.373	0.706	0.997	0.759	0.526
baseline_delete	0.739	0.808	0.75	0.461	0.739	1.0	0.682	0.499
nikita.sushko	0.764	0.798	0.693	0.437	0.764	1.0	0.646	0.491
baseline_gpt4	0.74	0.78	0.699	0.412	0.74	0.995	0.662	0.482
baseline_backtranslation	0.545	0.558	0.838	0.265	0.545	0.762	0.99	0.425
baseline_o3mini	0.554	0.575	0.897	0.291	0.554	0.828	0.916	0.421
SVATS	0.76	0.81	0.73	0.461	0.76	1.0	0.5	0.38
baseline_duplicate	0.76	0.81	0.73	0.461	0.76	1.0	0.5	0.38
SomethingAwful	0.468	0.494	0.814	0.205	0.468	0.807	0.945	0.367

Table 10
Automatic and LLM evaluation results for Arabic.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
Human References	0.913	0.95	1.0	0.869	0.913	0.945	0.969	0.838
Team MetaDetox	0.874	0.898	0.922	0.732	0.874	0.985	0.958	0.826
ducanhhbtt	0.869	0.869	0.941	0.718	0.869	0.975	0.961	0.814
Team Pratham	0.872	0.902	0.913	0.724	0.872	0.982	0.925	0.793
baseline_mt0	0.866	0.891	0.916	0.715	0.866	0.973	0.937	0.791
adugeen	0.872	0.9	0.897	0.713	0.872	0.978	0.926	0.79
jellyproll	0.866	0.891	0.916	0.715	0.866	0.972	0.935	0.788
sky.Duan	0.866	0.891	0.916	0.715	0.866	0.968	0.937	0.787
Jiaozipi	0.841	0.83	0.963	0.682	0.841	0.932	0.985	0.779
humairafaridq	0.852	0.799	0.902	0.621	0.852	0.935	0.957	0.768
d1n910	0.816	0.735	0.967	0.589	0.816	0.885	0.982	0.72
ylmmcl	0.836	0.838	0.888	0.643	0.836	0.935	0.903	0.718
Gopal	0.831	0.84	0.915	0.648	0.831	0.897	0.951	0.718
SVATS	0.865	0.907	0.838	0.668	0.865	0.99	0.827	0.705
baseline_gpt4	0.828	0.756	0.946	0.603	0.828	0.85	0.956	0.686
nikita.sushko	0.823	0.827	0.813	0.612	0.823	0.88	0.865	0.652
baseline_delete	0.842	0.916	0.788	0.611	0.842	0.972	0.761	0.61
baseline_o3mini	0.77	0.633	0.987	0.498	0.77	0.747	0.986	0.595
SomethingAwful	0.768	0.65	0.979	0.508	0.768	0.737	0.989	0.592
baseline_duplicate	0.89	0.926	0.68	0.564	0.89	1.0	0.501	0.446
baseline_backtranslation	0.72	0.673	0.867	0.438	0.72	0.58	0.968	0.442

Table 11
Automatic and LLM evaluation results for German.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
Human References	0.971	0.964	1.0	0.936	0.971	0.978	0.979	0.93
Team MetaDetox	0.954	0.922	0.868	0.766	0.954	0.99	0.973	0.919
ducanhhbtt	0.95	0.914	0.917	0.798	0.95	0.987	0.981	0.919
humairafaridq	0.944	0.89	0.866	0.729	0.944	0.987	0.978	0.912
adugeen	0.951	0.934	0.862	0.769	0.951	0.987	0.945	0.888
d1n910	0.909	0.805	0.94	0.693	0.909	0.955	0.99	0.866
SVATS	0.93	0.904	0.891	0.754	0.93	0.94	0.974	0.854
Jiaozipi	0.918	0.851	0.951	0.748	0.918	0.917	0.987	0.84
sky.Duan	0.949	0.925	0.859	0.757	0.949	0.968	0.905	0.83
nikita.sushko	0.938	0.907	0.853	0.727	0.938	0.957	0.925	0.828
jellyproll	0.949	0.925	0.859	0.757	0.949	0.968	0.902	0.828
baseline_mt0	0.949	0.925	0.859	0.757	0.949	0.962	0.906	0.825
Team Pratham	0.951	0.933	0.843	0.75	0.951	0.977	0.888	0.822
Gopal	0.931	0.897	0.84	0.707	0.931	0.937	0.938	0.819
baseline_gpt4	0.914	0.826	0.957	0.728	0.914	0.888	0.979	0.807
ylmmcl	0.895	0.856	0.861	0.684	0.895	0.888	0.933	0.772
SomethingAwful	0.844	0.736	0.972	0.618	0.844	0.853	0.998	0.749
baseline_o3mini	0.847	0.731	0.964	0.607	0.847	0.88	0.987	0.747
baseline_delete	0.949	0.941	0.657	0.586	0.949	0.985	0.612	0.564
baseline_backtranslation	0.81	0.741	0.838	0.513	0.81	0.558	0.975	0.479
baseline_duplicate	0.959	0.946	0.63	0.572	0.959	1.0	0.5	0.479

Table 12
Automatic and LLM evaluation results for English.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
baseline_o3mini	0.904	0.768	0.981	0.688	0.904	0.995	0.992	0.893
Team MetaDetox	0.915	0.861	0.933	0.742	0.915	0.995	0.981	0.893
humairafaridq	0.904	0.848	0.941	0.726	0.904	0.995	0.988	0.888
Jiaozipi	0.903	0.825	0.964	0.724	0.903	0.982	0.992	0.882
ducanhhbtt	0.893	0.861	0.948	0.734	0.893	0.993	0.982	0.871
baseline_gpt4	0.892	0.802	0.982	0.708	0.892	0.963	0.993	0.858
nikita.sushko	0.909	0.865	0.9	0.716	0.909	0.972	0.969	0.858
SomethingAwful	0.883	0.735	0.984	0.647	0.883	0.97	0.996	0.856
sky.Duan	0.893	0.872	0.92	0.727	0.893	0.967	0.977	0.847
Human References	0.885	0.928	1.0	0.822	0.885	0.96	0.993	0.846
baseline_mt0	0.893	0.872	0.92	0.727	0.893	0.965	0.976	0.843
Team Pratham	0.896	0.875	0.918	0.729	0.896	0.965	0.972	0.843
jellyproll	0.893	0.872	0.92	0.727	0.893	0.963	0.975	0.842
SVATS	0.89	0.842	0.929	0.704	0.89	0.96	0.968	0.83
d1n910	0.837	0.773	0.964	0.632	0.837	0.96	0.991	0.805
ylmmcl	0.859	0.82	0.922	0.667	0.859	0.92	0.984	0.796
adugeen	0.836	0.865	0.935	0.685	0.836	0.967	0.978	0.794
baseline_backtranslation	0.862	0.851	0.918	0.684	0.862	0.867	0.97	0.743
Gopal	0.743	0.85	0.917	0.583	0.743	0.932	0.99	0.691
baseline_delete	0.663	0.882	0.826	0.473	0.663	0.768	0.89	0.453
baseline_duplicate	0.739	0.892	0.533	0.353	0.739	1.0	0.5	0.37

Table 13
Automatic and LLM evaluation results for Spanish.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
Team MetaDetox	0.902	0.85	0.928	0.719	0.902	0.962	0.948	0.823
humairafaridq	0.896	0.83	0.915	0.688	0.896	0.95	0.956	0.814
Jiaozipi	0.893	0.817	0.966	0.712	0.893	0.927	0.974	0.812
d1n910	0.895	0.805	0.925	0.673	0.895	0.938	0.959	0.807
baseline_gpt4	0.901	0.814	0.958	0.708	0.901	0.92	0.96	0.8
ducanhhbtt	0.889	0.831	0.919	0.686	0.889	0.925	0.963	0.797
baseline_o3mini	0.897	0.762	0.955	0.66	0.897	0.912	0.968	0.796
Human References	0.933	0.924	1.0	0.863	0.933	0.87	0.964	0.783
adugeen	0.883	0.862	0.919	0.709	0.883	0.932	0.928	0.765
baseline_mt0	0.879	0.86	0.909	0.696	0.879	0.935	0.928	0.764
Team Pratham	0.88	0.865	0.901	0.696	0.88	0.947	0.917	0.763
SomethingAwful	0.878	0.764	0.959	0.652	0.878	0.882	0.974	0.763
jellyproll	0.879	0.86	0.909	0.696	0.879	0.927	0.929	0.758
sky.Duan	0.879	0.86	0.909	0.696	0.879	0.927	0.928	0.757
Gopal	0.871	0.842	0.892	0.662	0.871	0.927	0.936	0.757
SVATS	0.89	0.835	0.929	0.698	0.89	0.92	0.914	0.749
nikita.sushko	0.878	0.843	0.888	0.666	0.878	0.922	0.912	0.74
ylmmcl	0.785	0.733	0.925	0.581	0.785	0.785	0.932	0.624
baseline_delete	0.854	0.878	0.795	0.603	0.854	0.882	0.744	0.543
baseline_backtranslation	0.799	0.748	0.862	0.528	0.799	0.605	0.934	0.466
baseline_duplicate	0.902	0.887	0.7	0.566	0.902	1.0	0.5	0.451

Table 14
Automatic and LLM evaluation results for Hindi.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
Human References	0.906	0.926	1.0	0.841	0.906	0.888	0.993	0.807
Team MetaDetox	0.853	0.857	0.843	0.629	0.853	0.942	0.969	0.785
adugeen	0.851	0.869	0.815	0.619	0.851	0.927	0.971	0.773
ducanhhbtt	0.853	0.85	0.838	0.619	0.853	0.912	0.972	0.762
jellyproll	0.844	0.856	0.85	0.627	0.844	0.915	0.975	0.759
Team Pratham	0.849	0.863	0.844	0.631	0.849	0.918	0.965	0.758
baseline_mt0	0.844	0.856	0.85	0.627	0.844	0.903	0.976	0.751
sky.Duan	0.844	0.856	0.85	0.627	0.844	0.897	0.976	0.747
nikita.sushko	0.849	0.855	0.795	0.594	0.849	0.928	0.916	0.732
humairafaridq	0.836	0.804	0.853	0.584	0.836	0.893	0.964	0.732
ylmmcl	0.817	0.82	0.833	0.578	0.817	0.865	0.973	0.716
Gopal	0.826	0.842	0.806	0.573	0.826	0.898	0.933	0.701
Jiaozipi	0.825	0.824	0.908	0.631	0.825	0.833	0.992	0.701
d1n910	0.816	0.774	0.904	0.583	0.816	0.83	0.989	0.689
SVATS	0.827	0.746	0.939	0.593	0.827	0.793	0.988	0.672
baseline_gpt4	0.823	0.782	0.922	0.605	0.823	0.785	0.968	0.647
SomethingAwful	0.8	0.737	0.958	0.58	0.8	0.76	0.993	0.631
baseline_o3mini	0.789	0.707	0.954	0.549	0.789	0.747	0.986	0.609
baseline_delete	0.848	0.886	0.63	0.48	0.848	0.983	0.689	0.566
baseline_duplicate	0.865	0.889	0.539	0.417	0.865	1.0	0.5	0.432
baseline_backtranslation	0.713	0.712	0.791	0.419	0.713	0.553	0.948	0.395

Table 15
Automatic and LLM evaluation results for Russian.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
nikita.sushko	0.894	0.872	0.83	0.657	0.894	0.963	0.968	0.835
Team MetaDetox	0.901	0.869	0.952	0.753	0.901	0.922	0.994	0.829
ducanhhbtt	0.892	0.857	0.972	0.749	0.892	0.925	0.998	0.827
jellyproll	0.895	0.882	0.944	0.754	0.895	0.92	0.988	0.818
Jiaozipi	0.893	0.836	0.969	0.73	0.893	0.91	0.998	0.817
Team Pratham	0.896	0.884	0.944	0.755	0.896	0.915	0.986	0.811
sky.Duan	0.896	0.883	0.943	0.754	0.896	0.913	0.988	0.811
baseline_mt0	0.895	0.882	0.944	0.754	0.895	0.91	0.988	0.809
humairafaridq	0.882	0.844	0.933	0.703	0.882	0.912	0.992	0.805
SVATS	0.885	0.878	0.921	0.725	0.885	0.908	0.988	0.798
adugeen	0.896	0.87	0.955	0.75	0.896	0.885	0.994	0.792
Gopal	0.87	0.856	0.88	0.665	0.87	0.91	0.988	0.792
d1n910	0.885	0.799	0.954	0.681	0.885	0.893	0.995	0.792
Human References	0.931	0.93	1.0	0.867	0.931	0.842	0.993	0.783
baseline_gpt4	0.884	0.816	0.97	0.706	0.884	0.88	0.989	0.778
ylmmcl	0.854	0.833	0.939	0.69	0.854	0.878	0.987	0.76
baseline_o3mini	0.861	0.739	0.987	0.638	0.861	0.808	0.997	0.711
baseline_backtranslation	0.871	0.859	0.919	0.696	0.871	0.782	0.987	0.689
SomethingAwful	0.848	0.732	0.991	0.627	0.848	0.788	0.998	0.685
baseline_delete	0.869	0.889	0.663	0.514	0.869	0.927	0.746	0.583
baseline_duplicate	0.899	0.895	0.525	0.424	0.899	1.0	0.5	0.45

Table 16
Automatic and LLM evaluation results for Ukrainian.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
Team MetaDetox	0.904	0.913	0.961	0.798	0.904	0.882	0.987	0.791
adugeen	0.903	0.913	0.964	0.8	0.903	0.89	0.98	0.791
ducanhhbtt	0.899	0.907	0.975	0.799	0.899	0.878	0.984	0.785
Team Pratham	0.898	0.917	0.937	0.776	0.898	0.895	0.968	0.782
Human References	0.936	0.96	1.0	0.899	0.936	0.843	0.984	0.78
nikita.sushko	0.899	0.909	0.896	0.738	0.899	0.907	0.944	0.772
sky.Duan	0.895	0.91	0.938	0.77	0.895	0.878	0.976	0.771
baseline_mt0	0.895	0.91	0.938	0.77	0.895	0.878	0.974	0.77
jellyproll	0.895	0.91	0.938	0.77	0.895	0.872	0.975	0.766
d1n910	0.879	0.851	0.973	0.733	0.879	0.855	0.993	0.757
Jiaozipi	0.887	0.881	0.982	0.773	0.887	0.847	0.993	0.757
humairafaridq	0.89	0.861	0.977	0.754	0.89	0.837	0.992	0.748
SVATS	0.885	0.915	0.938	0.766	0.885	0.855	0.972	0.743
Gopal	0.884	0.88	0.923	0.725	0.884	0.853	0.974	0.742
ylmmcl	0.85	0.845	0.934	0.693	0.85	0.853	0.974	0.727
baseline_gpt4	0.886	0.847	0.987	0.747	0.886	0.81	0.991	0.723
SomethingAwful	0.859	0.803	0.985	0.689	0.859	0.793	0.993	0.694
baseline_o3mini	0.867	0.785	0.994	0.685	0.867	0.748	0.995	0.663
baseline_delete	0.884	0.932	0.707	0.581	0.884	0.903	0.747	0.577
baseline_duplicate	0.91	0.94	0.516	0.442	0.91	1.0	0.5	0.455
baseline_backtranslation	0.708	0.708	0.959	0.498	0.708	0.308	0.992	0.256

Table 17
Automatic and LLM evaluation results for Chinese.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
Human References	0.927	0.916	1.0	0.85	0.927	0.97	0.986	0.887
Team MetaDetox	0.844	0.804	0.892	0.611	0.844	0.978	0.981	0.813
ducanhhbtt	0.836	0.803	0.91	0.618	0.836	0.982	0.968	0.796
adugeen	0.85	0.851	0.732	0.543	0.85	0.987	0.933	0.783
SVATS	0.852	0.855	0.707	0.531	0.852	0.985	0.925	0.776
Jiaozipi	0.792	0.698	0.958	0.539	0.792	0.93	0.992	0.744
ylmmcl	0.843	0.818	0.744	0.523	0.843	0.932	0.908	0.725
humairafaridq	0.82	0.765	0.858	0.546	0.82	0.978	0.901	0.725
baseline_mt0	0.844	0.843	0.752	0.543	0.844	0.967	0.88	0.717
Team Pratham	0.805	0.83	0.79	0.533	0.805	0.94	0.942	0.717
jellyproll	0.844	0.843	0.752	0.543	0.844	0.96	0.884	0.715
nikita.sushko	0.852	0.843	0.689	0.509	0.852	0.97	0.848	0.702
Gopal	0.797	0.792	0.889	0.57	0.797	0.918	0.945	0.699
sky.Duan	0.841	0.831	0.784	0.551	0.841	0.988	0.838	0.696
baseline_gpt4	0.783	0.709	0.912	0.513	0.783	0.917	0.902	0.654
baseline_o3mini	0.734	0.599	0.977	0.439	0.734	0.922	0.952	0.652
d1n910	0.752	0.618	0.967	0.461	0.752	0.872	0.975	0.652
baseline_delete	0.748	0.818	0.831	0.516	0.748	0.902	0.926	0.63
SomethingAwful	0.718	0.571	0.985	0.418	0.718	0.858	0.985	0.629
baseline_duplicate	0.857	0.874	0.63	0.477	0.857	1.0	0.501	0.429
baseline_backtranslation	0.593	0.54	0.887	0.29	0.593	0.355	0.965	0.231

Table 18
Automatic and LLM evaluation results for French.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
ducanhhbtt	0.926	0.918	0.939	0.802	0.926	0.965	0.992	0.889
Human References	0.941	0.961	1.0	0.905	0.941	0.945	0.997	0.888
Team MetaDetox	0.928	0.904	0.951	0.802	0.928	0.952	0.995	0.883
humairafaridq	0.918	0.893	0.897	0.74	0.918	0.963	0.995	0.883
sky.Duan	0.921	0.898	0.841	0.699	0.921	0.962	0.982	0.873
baseline_gpt4	0.921	0.871	0.969	0.78	0.921	0.938	0.996	0.865
adugeen	0.918	0.924	0.922	0.785	0.918	0.95	0.983	0.86
d1n910	0.915	0.848	0.955	0.744	0.915	0.932	0.994	0.853
Jiaozipi	0.922	0.884	0.98	0.801	0.922	0.92	0.998	0.85
Gopal	0.899	0.907	0.853	0.7	0.899	0.953	0.979	0.843
baseline_o3mini	0.905	0.802	0.991	0.725	0.905	0.908	0.997	0.826
SVATS	0.91	0.884	0.954	0.769	0.91	0.902	0.986	0.815
SomethingAwful	0.899	0.804	0.984	0.717	0.899	0.895	0.998	0.814
nikita.sushko	0.908	0.893	0.926	0.754	0.908	0.892	0.988	0.807
Team Pratham	0.908	0.914	0.902	0.752	0.908	0.902	0.973	0.801
baseline_mt0	0.908	0.911	0.915	0.76	0.908	0.885	0.978	0.793
jellyproll	0.908	0.911	0.915	0.76	0.908	0.883	0.978	0.79
ylmmcl	0.82	0.8	0.905	0.625	0.82	0.772	0.969	0.659
baseline_delete	0.859	0.941	0.645	0.518	0.859	0.91	0.762	0.576
baseline_backtranslation	0.82	0.8	0.94	0.626	0.82	0.573	0.992	0.503
baseline_duplicate	0.92	0.942	0.516	0.447	0.92	1.0	0.5	0.46

Table 19
Automatic and LLM evaluation results for Hebrew.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
ducanhhbtt	0.798	0.739	0.873	0.531	0.798	0.908	0.938	0.681
humairafaridq	0.787	0.727	0.829	0.489	0.787	0.888	0.947	0.671
adugeen	0.772	0.739	0.805	0.479	0.772	0.923	0.912	0.657
Gopal	0.758	0.73	0.831	0.478	0.758	0.883	0.932	0.631
sky.Duan	0.781	0.732	0.754	0.446	0.781	0.913	0.869	0.619
Jiaozipi	0.754	0.705	0.903	0.499	0.754	0.798	0.983	0.611
Team MetaDetox	0.795	0.707	0.917	0.53	0.795	0.77	0.97	0.61
d1n910	0.773	0.681	0.934	0.507	0.773	0.743	0.984	0.581
baseline_gpt4	0.769	0.7	0.922	0.513	0.769	0.75	0.966	0.578
SVATS	0.783	0.729	0.76	0.451	0.783	0.925	0.803	0.576
nikita.sushko	0.786	0.726	0.757	0.449	0.786	0.88	0.837	0.573
Human References	0.907	0.848	1.0	0.772	0.907	0.58	0.986	0.528
baseline_mt0	0.705	0.696	0.81	0.415	0.705	0.782	0.902	0.501
baseline_o3mini	0.738	0.637	0.974	0.475	0.738	0.647	0.985	0.497
baseline_delete	0.772	0.764	0.715	0.436	0.772	0.977	0.678	0.496
Team Pratham	0.697	0.693	0.828	0.416	0.697	0.763	0.918	0.495
jellyproll	0.705	0.696	0.81	0.415	0.705	0.777	0.901	0.495
SomethingAwful	0.739	0.643	0.96	0.473	0.739	0.627	0.992	0.49
baseline_duplicate	0.814	0.772	0.663	0.425	0.814	1.0	0.5	0.407
ylmmcl	0.601	0.557	0.886	0.323	0.601	0.57	0.94	0.357
baseline_backtranslation	0.652	0.63	0.786	0.339	0.652	0.538	0.938	0.349

Table 20
Automatic and LLM evaluation results for Hinglish.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
Human References	0.84	0.927	1.0	0.781	0.84	0.86	0.982	0.716
Team MetaDetox	0.733	0.762	0.843	0.481	0.733	0.86	0.968	0.621
adugeen	0.743	0.832	0.771	0.491	0.743	0.917	0.888	0.606
ducanhhbtt	0.739	0.786	0.863	0.511	0.739	0.853	0.933	0.592
humairafaridq	0.733	0.721	0.781	0.422	0.733	0.865	0.914	0.587
d1n910	0.702	0.67	0.865	0.416	0.702	0.835	0.974	0.584
Gopal	0.732	0.845	0.745	0.47	0.732	0.922	0.867	0.578
SVATS	0.742	0.861	0.689	0.455	0.742	0.963	0.777	0.549
Jiaozipi	0.707	0.747	0.889	0.48	0.707	0.772	0.968	0.544
baseline_gpt4	0.754	0.53	0.816	0.333	0.754	0.75	0.919	0.524
sky.Duan	0.741	0.398	0.648	0.196	0.741	0.783	0.879	0.509
jellyproll	0.746	0.885	0.67	0.449	0.746	0.978	0.702	0.503
Team Pratham	0.687	0.773	0.659	0.356	0.687	0.878	0.85	0.502
baseline_mt0	0.697	0.791	0.621	0.351	0.697	0.925	0.783	0.494
baseline_delete	0.748	0.885	0.631	0.425	0.748	0.987	0.67	0.486
SomethingAwful	0.669	0.619	0.92	0.392	0.669	0.695	0.985	0.477
baseline_o3mini	0.679	0.399	0.906	0.251	0.679	0.603	0.968	0.411
nikita.sushko	0.689	0.607	0.586	0.262	0.689	0.728	0.803	0.41
baseline_duplicate	0.773	0.89	0.601	0.419	0.773	1.0	0.5	0.387
ylmmcl	0.551	0.488	0.775	0.213	0.551	0.563	0.904	0.312
baseline_backtranslation	0.427	0.323	0.909	0.133	0.427	0.308	0.988	0.139

Table 21
Automatic and LLM evaluation results for Italian.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
Human References	0.94	0.976	1.0	0.918	0.94	0.973	0.976	0.893
ducanhhbtt	0.883	0.9	0.982	0.784	0.883	0.963	0.987	0.842
adugeen	0.879	0.92	0.938	0.761	0.879	0.958	0.975	0.823
sky.Duan	0.88	0.848	0.88	0.663	0.88	0.955	nan	0.822
Team MetaDetox	0.883	0.87	0.977	0.755	0.883	0.932	0.992	0.821
humairafaridq	0.879	0.833	0.948	0.7	0.879	0.94	0.985	0.819
Gopal	0.871	0.915	0.933	0.746	0.871	0.953	0.975	0.812
Jiaozipi	0.863	0.851	0.983	0.728	0.863	0.912	0.998	0.795
d1n910	0.874	0.834	0.95	0.698	0.874	0.913	0.987	0.791
baseline_gpt4	0.876	0.851	0.99	0.742	0.876	0.907	0.988	0.79
SVATS	0.883	0.9	0.948	0.755	0.883	0.967	0.909	0.775
nikita.sushko	0.869	0.902	0.923	0.727	0.869	0.945	0.927	0.763
Team Pratham	0.867	0.927	0.929	0.749	0.867	0.955	0.91	0.752
baseline_mt0	0.864	0.92	0.932	0.746	0.864	0.945	0.918	0.749
baseline_o3mini	0.843	0.722	0.975	0.605	0.843	0.875	0.995	0.748
jellyproll	0.864	0.92	0.932	0.746	0.864	0.938	0.916	0.742
SomethingAwful	0.839	0.765	0.978	0.637	0.839	0.848	0.998	0.728
ylmmcl	0.8	0.834	0.923	0.645	0.8	0.857	0.928	0.662
baseline_delete	0.835	0.95	0.844	0.668	0.835	0.917	0.835	0.628
baseline_duplicate	0.91	0.964	0.744	0.653	0.91	1.0	0.5	0.455
baseline_backtranslation	0.707	0.696	0.915	0.462	0.707	0.438	0.982	0.333

Table 22
Automatic and LLM evaluation results for Japanese.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
Human References	0.92	0.961	1.0	0.884	0.92	0.988	0.993	0.904
ducanhhbtt	0.851	0.893	0.881	0.674	0.851	0.993	0.969	0.82
humairafaridq	0.854	0.858	0.898	0.663	0.854	0.98	0.974	0.819
adugeen	0.868	0.919	0.795	0.64	0.868	0.995	0.932	0.805
d1n910	0.828	0.826	0.917	0.637	0.828	0.965	0.99	0.796
Jiaozipi	0.813	0.826	0.953	0.647	0.813	0.962	0.996	0.787
Gopal	0.837	0.888	0.884	0.662	0.837	0.967	0.965	0.784
baseline_gpt4	0.832	0.821	0.926	0.637	0.832	0.965	0.964	0.779
sky.Duan	0.842	0.856	0.79	0.573	0.842	0.975	0.934	0.769
jellyproll	0.777	0.909	0.905	0.644	0.777	0.958	0.991	0.745
SVATS	0.86	0.9	0.761	0.589	0.86	0.995	0.859	0.734
nikita.sushko	0.829	0.777	0.747	0.495	0.829	0.953	0.906	0.722
Team MetaDetox	0.784	0.76	0.968	0.587	0.784	0.908	0.992	0.721
baseline_mt0	0.855	0.917	0.74	0.582	0.855	0.99	0.842	0.711
Team Pratham	0.86	0.923	0.741	0.591	0.86	0.992	0.834	0.71
baseline_o3mini	0.745	0.689	0.937	0.49	0.745	0.88	0.986	0.661
ylmmcl	0.829	0.87	0.722	0.528	0.829	0.933	0.819	0.647
SomethingAwful	0.723	0.699	0.951	0.496	0.724	0.858	0.996	0.643
baseline_delete	0.884	0.941	0.531	0.441	0.884	1.0	0.501	0.443
baseline_duplicate	0.884	0.941	0.53	0.44	0.884	1.0	0.5	0.442
baseline_backtranslation	0.493	0.519	0.899	0.241	0.493	0.243	0.998	0.147

Table 23
Automatic and LLM evaluation results for Tatar.

	Automatic Evaluation				LLM evaluation			
	FL	SIM	STA	J	FL	SIM	STA	J*
Human References	0.878	0.936	1.0	0.825	0.878	0.852	0.958	0.724
jellyproll	0.803	0.869	0.868	0.617	0.803	0.932	0.826	0.611
baseline_mt0	0.827	0.879	0.779	0.58	0.827	0.945	0.774	0.598
adugeen	0.799	0.824	0.86	0.582	0.799	0.863	0.848	0.583
Team Pratham	0.83	0.883	0.78	0.584	0.83	0.948	0.74	0.575
Gopal	0.792	0.807	0.789	0.516	0.792	0.877	0.828	0.575
ylmmcl	0.764	0.755	0.8	0.492	0.764	0.853	0.818	0.543
Jiaozipi	0.758	0.762	0.811	0.485	0.758	0.772	0.909	0.541
SVATS	0.826	0.899	0.764	0.573	0.826	0.992	0.647	0.523
baseline_delete	0.826	0.899	0.764	0.573	0.826	0.987	0.648	0.521
humairafaridq	0.774	0.733	0.779	0.452	0.774	0.758	0.861	0.511
ducanhhbtt	0.757	0.78	0.923	0.556	0.757	0.707	0.908	0.495
Team MetaDetox	0.796	0.732	0.842	0.498	0.796	0.712	0.882	0.493
nikita.sushko	0.804	0.762	0.622	0.386	0.804	0.837	0.758	0.492
baseline_gpt4	0.796	0.739	0.789	0.468	0.796	0.672	0.848	0.438
d1n910	0.74	0.693	0.85	0.45	0.74	0.635	0.905	0.436
baseline_duplicate	0.842	0.904	0.661	0.51	0.842	1.0	0.5	0.421
sky.Duan	0.757	0.73	0.763	0.427	0.757	0.683	0.834	0.416
SomethingAwful	0.658	0.604	0.847	0.352	0.658	0.468	0.954	0.324
baseline_o3mini	0.718	0.556	0.873	0.36	0.718	0.293	0.951	0.209
baseline_backtranslation	0.567	0.491	0.889	0.254	0.567	0.085	0.982	0.054

B. Prompts used for LLM as a judge

`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:

`Flyency 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:

`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: