GemDetox at TextDetox CLEF 2025: Enhancing a Massively Multilingual Model for Text Detoxification on Low-resource Languages

Notebook for the PAN Lab at CLEF 2025

Trung Duc Anh Dang[1], Ferdinando Pio D'Elia[1,*]

[1]Centre for Language Technology, University of Copenhagen, Denmark

Abstract

As social-media platforms emerge and evolve faster than the regulations meant to oversee them, automated detoxification might serve as a timely tool for moderators to enforce safe discourse at scale. We here describe our submission to the PAN 2025 Multilingual Text Detoxification Challenge, which rewrites toxic single-sentence inputs into neutral paraphrases across 15 typologically diverse languages. Building on a 12B-parameter Gemma-3 multilingual transformer, we apply parameter-efficient LoRA SFT fine-tuning and prompting techniques like few-shot and Chain-of-Thought. Our multilingual training corpus combines 3 600 human-authored parallel pairs, 21 600 machine-translated synthetic pairs, and model-generated pairs filtered by Jaccard thresholds. At inference, inputs are enriched with three LaBSE-retrieved neighbors and explicit toxic-span annotations. Evaluated via Style Transfer Accuracy, LaBSE-based semantic preservation, and xCOMET fluency, our system ranks first on high-resource and low-resource languages. Ablations show +0.081 joint score increase from few-shot examples and +0.088 from basic CoT prompting. ANOVA analysis identifies language resource status as the strongest predictor of performance ($\eta^2 = 0.667$, $p < 0.01$).

Warning: This paper contains offensive and potentially triggering texts that only serve as illustrative examples.

Keywords

Multilingual text detoxification, Large Language Models, Parameter-efficient fine-tuning, Chain-of-Thought, Data augmentation

1. Introduction

The widespread use of digital communication and social media platforms has prompted the need, rather urgent, for moderation and content detoxification strategies that can be proved effective and easy to implement. Toxic language, hate speech, and harassment jeopardize safety and pluralism of online spaces, motivating the research community to fortify automated methods that could intervene at scale. The PAN 2025 Multilingual Detoxification Challenge [1][2] offers a shared testbed for systems that aim to rephrase user posts into safer language while keeping their original meaning, and across multiple languages.

Large-scale detoxification is not a purely technical exercise; it carries substantial social, political, and practical ramifications. In real-world scenarios, excessive detoxification interventions may be perceived as censorship, eroding trust and discouraging open dialogue; interventions may strip away the original message of its core purpose. The stakes are even higher for minorities and marginalized communities, who nowadays often rely on these platforms to voice political and social dissent and might consequently feel censored: as far as dissenting voices are concerned, detoxification could silence the very people it aims to protect. While we set these broader implications aside for the present study, they remain a key source of motivation for our work.

The Multilingual Text Detoxification task at PAN 2025 challenges participants to rewrite a toxic piece of text into a non-toxic form while preserving as much of the original meaning as possible across 15

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

cls364@alumni.ku.dk (T. D. A. Dang); plh618@alumni.ku.dk (F. P. D'Elia)

https://github.com/ducanhdt/ducanhhbtt-detoxification2025/ (T. D. A. Dang)

typologically diverse languages, ranging from English and Spanish to Tatar and Hinglish. Formally, the input is a single-sentence text that contains at least one instance of toxic language, and the system must produce a semantically faithful paraphrase with neutral tone. Notably, the competition constrains the notion of toxicity to explicit toxicity: obscene or offensive lexicon in which meaningful neutral content is still present. Implicit toxicity, such as sarcasm, coded hate speech, or passive-aggressive formulations, is excluded[1]. We here present our approach, which builds directly on the 2024 paradigm[4] of fine-tuning large-scale multilingual pretrained transformer models for detoxification. We opted for a model with extensive multilingual pre-training yet lightweight features, and focused on maximizing the possibilities of Chain-of-Thought prompting as well as data augmentation.

The paper is organized as follows: in Section 2 we review related work; Section 3 describes our methodology; Section 4 details the experimental setup; Section 5 presents our automatic and qualitative results; and Section 6 discusses limitations and future directions. Our experiments show that, while our approach achieves strong performance across all languages, the gains from data augmentation are especially pronounced for the low-resource languages for which they were designed, narrowing their gap with high-resource counterparts.

2. Related work

Text detoxification, a specialized subfield of Text Style Transfer (TST), involves transforming toxic texts into neutral versions while maintaining semantic integrity and linguistic fluency. Initial research in detoxification was largely driven by prominent competitions such as the Jigsaw/Conversation AI Kaggle challenges (2018–2021) [5], providing substantial datasets and significantly advancing toxicity detection methods. Early models, such as [6], employed unsupervised encoder-decoder architectures with cycle consistency losses for style transfer. Subsequent approaches, such as CondBERT and ParaGedi [7], introduced unsupervised conditional masked language modeling and paraphrasing, setting benchmarks at the time. Further progress was marked by the ParaDetox corpus [8], which demonstrated that parallel data substantially improved the detoxification performance over purely unsupervised techniques.

Recently, transformer-based large language models (LLMs) have been largely utilized in detoxification research due to their powerful text generation capabilities. Notably, GPT-based models, including GPT-2, GPT-3, and GPT-4, demonstrated considerable efficacy when fine-tuned or used with few-shot prompting strategies [9]. Models such as GPT-DETOX have explored innovative in-context learning techniques including zero-shot, few-shot, and ensemble prompting, significantly outperforming earlier supervised and unsupervised methods.

Parallel corpus availability remains critical for enhancing detoxification methods. The creation of new datasets such as the multilingual MultiParaDetox corpus [10] has significantly expanded detoxification research to include languages previously underrepresented, such as Hindi, Arabic, Chinese, and Amharic. Furthermore, sophisticated prompt-engineering frameworks, like CO-STAR [11], have shown promising results by strategically guiding LLMs to enhance contextual and semantic coherence during detoxification. Recent advancements have also emphasized explainability and interpretability in detoxification processes. Very recently, a first automated explainable analysis across multiple languages was published [12], revealing common patterns and language-specific toxicity traits, and implemented Chain-of-Thought reasoning techniques to enhance the detoxification accuracy of LLMs. Hybrid detoxification models combining editing-based and sequence-to-sequence approaches, as demonstrated in [13] with DiffuDetox, leverage the complementary strengths of multiple techniques. These hybrid approaches have achieved state-of-the-art performance in automatic evaluations and showed good results in human assessments compared to purely editing-based or sequence-to-sequence models.

Overall, text detoxification research now involves a broad spectrum of methodologies from unsu-

[1]It warrants mention that such subtle toxicity is just as dangerous as a social phenomenon, and ample research has shown how ordinary terms and euphemisms can smuggle dehumanizing ideas into everyday speech, gradually normalizing hatred and exclusion. [3] Yet, any problem is such only when we can carve it into a series of smaller and more manageable steps; this is our starting point.

pervised conditional generation and fine-tuned LLMs to explainable AI and advanced multilingual prompting strategies. Competitions such as the Multilingual Text Detoxification Task at PAN 2024 offer valuable opportunities to explore the capabilities and limitations of LLMs, particularly in their lightweight versions.

3. Methods

3.1. Datasets

This section details the construction of the training corpus we employed in all experiments. The final corpus merges (i) the organiser-provided parallel data, (ii) machine-translated extensions for six missing languages, and (iii) synthetic pairs mined from a toxicity-only collection. The PAN 2025 Multilingual Text-Detoxification shared task [2] supplies a parallel corpus (ParaDetox-9). It comprises 400 toxic–neutral sentence pairs for each of nine languages: English, Spanish, German, Chinese, Arabic, Hindi, Ukrainian, Russian and Amharic, for a total of 3 600 pairs. These data constitute the only human-authored supervision available for the task.

3.1.1. Data augmentation for unseen languages

The remaining six target languages (Italian, French, Hebrew, Hinglish, Japanese and Tatar) are not covered by ParaDetox-9. To extend coverage to these languages, we translated all the 3 600 original pairs (in the 9 different "seen" languages) with two publicly available neural machine-translation systems: the RLM-Hinglish Translator [14] for Hinglish, and the NLLB-200 (3.3 B-parameter) [15] for Italian, French, Hebrew, Japanese and Tatar.

In all cases, each original toxic–neutral pair (s_1, s_2) from ParaDetox-9 was translated sentence-by-sentence into (t_1, t_2). The whole corpus of 3 600 source pairs was translated for each of the six unseen languages, thereby obtaining 21 600 pairs in total. This resulting corpus, named ParaDetox-MT6, is publicly released alongside our code and models [16].

3.1.2. Additional synthetic data generation

We incorporated the Multilingual Toxicity Dataset [17], which contains toxic sentences in 15 languages but no corresponding non-toxic rewrites. For each toxic sentence we generated one or more neutralization candidates with our strongest detoxification model (see further, 3.3) and retained only those pairs that satisfied the filtering criteria described below, 3.1.3. This procedure contributed additional supervision[2].

3.1.3. Filtering and cleaning

To ensure high-quality supervision, all candidate pairs, both machine-translated and model-generated, are processed by a uniform cleaning pipeline:

1. Jaccard similarity filtering, where we compute the character-level Jaccard index $J(s_{\text{toxic}}, s_{\text{clean}})$ over 5-grams and discard pairs with $J \geq 0.90$ to enforce lexical divergence.
2. Removal duplication, where pairs whose toxic and clean sentences are identical (after Unicode canonicalisation and lower-casing) are removed.
3. Hinglish script filtering, where candidate Hinglish language pairs containing any Devanagari code point are eliminated to avoid mixed-script noise[3].

[2] It is worth noting that recent work has already proposed interesting and useful synthetic data generation pipelines for multilingual detoxification objectives; see [18] for further details.

[3] In social and conversational text, Hinglish is almost exclusively written in the Latin (Roman) alphabet, with English words and romanised Hindi freely mixed. Any Devanagari code points usually signal a switch to standard-script Hindi, introducing (i) vocabulary sparsity, because the Hinglish tokenizer is trained on Roman script, and (ii) label ambiguity by blurring the boundary between the Hinglish and Hindi tracks.

Table 1
Toxic–neutral sentence pairs before and after the cleaning pipeline.

Language	Pairs before	Pairs after
English (en)	1 022	1 012
Spanish (es)	1 363	1 362
German (de)	1 185	1 142
Chinese (zh)	770	770
Arabic (ar)	1 022	1 019
Hindi (hi)	568	568
Ukrainian (uk)	2 535	2 535
Russian (ru)	1 565	1 520
Amharic (am)	719	719
Italian (it)	2 694	1 948
French (fr)	2 375	1 666
Hebrew (he)	1 772	1 037
Hinglish (hin)	1 135	561
Japanese (ja)	983	661
Tatar (tt)	1 951	1 275
Total	**21 659**	**17 795**

4. Semantic preservation filter, where a quality check is introduced on the pipeline to ensure the quality of the candidate neutral rewrites. To achieve this, pairs from the machine-translated augmentation set are retained only if $J>0.85$, while pairs synthesised from toxicity-only collections must satisfy the slightly looser $J>0.80$ threshold.

Lastly, to provide additional context during both training and evaluation, each toxic sentence is first embedded with LaBSE and associated with its three closest semantic neighbours in the same language. The identifiers of those six neighbours are stored alongside the sentence, giving downstream models quick access to in-language and cross-language examples that convey a similar meaning. A second enrichment pass applies a rule-based detector that extracts every explicit slur or profanity token and records them next to the sentence. The rule-based detector implements the same exact strategy of the Delete baseline provided for the PAN shared task.

3.1.4. Corpus statistics

Table 1 summarizes the number of retained pairs per language after cleaning. Altogether, the final corpus comprises ≈18 000 toxic–neutral pairs covering all 15 languages included in the PAN shared task.

3.2. Baselines

We evaluate our method against the unsupervised baselines provided by the PAN shared task. The trivial Duplicate baseline simply echoes the toxic source sentence unchanged. The Delete baseline removes any term appearing in the multilingual toxic lexicon [19] for each language without further rewriting. The Backtranslation baseline performs a two-step cross–lingual transfer: the input is first translated into English using the distilled NLLB-200 600M-parameter model [20], then detoxified with the English BART-base-detox [21] checkpoint, and finally translated back into the original language via NLLB-200 600M. Finally, zero-shot prompting baselines were also provided by the shared task committee: LLaMA-3.1-70B-Instruct-lorablated [22], as well as OpenAI's GPT-4-0613, GPT-4o-2024-08-06, and o3-mini-2025-01-31 [23].

3.3. Proposed model

As our base model we adopt the Gemma-3-12B-Instruct architecture[4] [24], quantised to 4-bit integer precision to cap memory usage below 24 GB while retaining the original 4 096-token rotary position scheme. At load time the Unsloth runtime converts activations to BF16 on Hopper-class GPUs and expands the context window to 4 028 tokens. To enable parameter-efficient adaptation we insert low-rank adapters (rank = 16, scaling = 16, no dropout) into every language, attention and MLP sub-module, freezing the remaining weights. This makes only 0.55% ($\approx$ 65M) of the 12B parameters trainable while preserving the expressiveness of the original network.

To better leverage the possibilities of the language model, we design a prompting strategy inspired by Chain-of-thought (CoT) principles, given the outstanding results these approaches have shown in recent years [25]. Namely, the system message outlines a four-step detoxification instruction which prompts the model to (i) identify toxic element(s), (ii) retrieve the semantic content of the overall sentence, (iii) rewrite using neutral words, (iv) check that your output be non-toxic. We created a base-prompt in English[5], and then passed it through OpenAI's o4-mini-high [23] on a sample of three languages in which we are proficient, either as native speakers or as second-language users: we manually checked them for consistency, and then proceeded to translate the base-prompt to the remaining languages using the same model. The result was 15 different prompts in the respective languages required by the PAN shared task.

Each training instance is therefore rendered as a three-turn process: the system message, the user supplying the toxic sentence with the language tag, the model's returned sentence pairs. For stronger supervision, the prompt prepends the semantically closest k examples (three in our experiments) drawn from the same language, retrieved from the datasets described in 3.1 by selecting the top three toxic sentences most similar in meaning, context, and phrasing to the target, yielding language-aware few-shot conditioning. We format the model's output in a standardized JSON structure. This improves output consistency, facilitates reliable parsing, and simplifies downstream processing for evaluation and training data collection.

During optimization, tokens belonging to the system and user turns are masked so that the cross-entropy objective is evaluated solely on the assistant span, preventing leakage of latent reasoning. Training data are further filtered by high semantic overlap ($J > 0.9$), inequality between toxic and neutral sentences, and membership in a seven-language target set. After fine-tuning, we convert the model into the VLLM format for low-latency, high-throughput inference. At inference time, we sample three candidate rewrites per toxic input and, using the available reference neutral sentence, compute the joint score J for each triple $\langle \text{toxic input}, \text{reference neutral}, \text{model output} \rangle$; the candidate with the highest J-score is chosen as the final detoxified output.

4. Experimental setup

4.1. Evaluation

For our experiments, we adhere to the evaluation protocol established by the shared task committee, which employs three primary metrics: Style Transfer Accuracy (STA), Content Preservation (SIM), and Fluency (FL)—each normalized to the interval [0,1].

Style Transfer Accuracy assesses both the absolute and relative non-toxicity of the generated output. Let p_{gen} denote the non-toxicity probability of the machine-generated sentence and $p_{\text{ref}}^{(i)}$ for $i = 1$ to

[4] In initial phases of our work, we tested other models as well: in quick prompt-only tests, the 7 B Qwen-2.5 and 27 B Gemma-3 variants detoxified more reliably than the 4 B editions, but the 12 B Gemma matched the 27 B's quality while fitting on a single 24 GB card. We therefore decided to adopt Gemma-3 especially because of its massively multilingual pre-training, which we believed could offer promising performance on zero-shot transferring. When presented to the public by the Google Dev team, Gemma3 was reported to be "available in 140 languages". Training data for the model was not made available by Google.

[5] The base prompt is available on our open-access repository on GitHub.

N the probabilities for each of the N human-authored detoxifications, as computed by a fine-tuned XLM-RoBERTa binary classifier [26]. We define:

$$\mathrm{STA} = \frac{p_{\mathrm{gen}} + \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\big(p_{\mathrm{gen}} \leq p_{\mathrm{ref}}^{(i)}\big)}{2}, \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function. This formulation penalizes outputs that are more toxic than human references while preventing over-rewarding outputs that simply match reference non-toxicity.

Content Preservation quantifies semantic fidelity through a weighted cosine similarity of LaBSE embeddings [27]:

$$\mathrm{SIM} = 0.4 \times \mathrm{CosSim}\big(\mathrm{input}, \mathrm{output}_{\mathrm{gen}}\big) + 0.6 \times \mathrm{CosSim}\big(\mathrm{output}_{\mathrm{gold}}, \mathrm{output}_{\mathrm{gen}}\big). \tag{2}$$

This balance ensures that the generated detoxification remains faithful both to the original toxic input and to the human-written reference. Lastly, Fluency is measured by the xCOMET[28] metric $X_{\mathrm{comet\_fluency}}$, which in experiments conducted by the shared task organizers, has demonstrated near-perfect correlation with human judgments of fluency in detoxified texts.

These metrics are combined into a single joint score J per sample:

$$J = X_{\mathrm{comet\_fluency}}\big(\mathrm{input}, \mathrm{output}_{\mathrm{gold}}, \mathrm{output}_{\mathrm{gen}}\big) \times \big(0.4 \times \mathrm{CosSim}(\mathrm{input}, \mathrm{output}_{\mathrm{gen}}) + 0.6 \times \mathrm{CosSim}(\mathrm{output}_{\mathrm{gold}}, \mathrm{output}_{\mathrm{gen}})\big) \times \mathrm{STA} \tag{3}$$

Here, the updated STA term incorporates human references by averaging the classifier's non-toxicity probability on the generated output with its relative position among the reference scores. This refinement penalizes outputs that remain more toxic than human-authored detoxifications, while the weighted similarity term balances fidelity to both the original input and the human detoxification.

4.2. Model training

We conduct Low-Rank Adaptation (LoRA) fine-tuning [29] on our multilingual corpus (with a 900-pair validation set) in three iterative phases. LoRA fine-tunes a large backbone by freezing its original weights W_0 and learning a low-rank update:

$$W = W_0 + \tfrac{\alpha}{r}\, AB, \qquad A \in \mathbb{R}^{d\times r},\ B \in \mathbb{R}^{r\times k}.$$

With $r = \alpha = 16$ on the 12-B-parameter Gemma-3-12B backbone this adds $\approx$ 65 M trainable parameters (0.55%). Combined with 4-bit quantisation and Unsloth's fused kernels, this keeps an 8-sample batch on a single NVIDIA A100 (80 GB) without gradient accumulation. Adapters are inserted in every self-attention projection and the MLP gating/up/down projections; dropout is 0, bias is *none*, seed = 3407. We train in three iterative phases:

- Phase 1. We fine-tune across all 15 languages for 1 000 optimisation steps with 8-bit AdamW (peak $LR = 2 \times 10^{-5}$; 20 warm-up steps; linear decay; weight decay = 0.01; seed = 3 407). This broadly aligns the model across high-resource (seen) and low-resource (unseen) languages.
- Phase 2. We apply a second fine-tuning pass using only data for the languages without available parallel data, running 1 000 steps at a reduced $LR = 5 \times 10^{-6}$ to avoid catastrophic interference, thereby preserving the first phase for high-resource languages. We save this adapter as Checkpoint II.
- Phase 3. We apply the best model from Phase 2 to generate non-toxic rewrites for toxic-only sentences in the synthetic dataset (described above in 3.1.2), filtered by Jaccard > 0.9, then fine-tune for a final 1 000 steps. We save this adapter as Checkpoint III.

At inference, languages without available parallel data are handled by the Checkpoint II, whereas languages without available parallel data are processed by the more specialized adapter from the third phase (Checkpoint III).

Across all phases (3 000 steps, $\tilde{3}$ hours total), memory is conserved via Unsloth's gradient checkpointing and smart off-loading; checkpoints are saved every 100 steps, and token-level perplexity (training loss) is logged to Weights & Biases at 20-step intervals. At inference, three candidate rewrites are sampled per input and the best one is chosen.

5. Results

Our final automatic results on the Test set are reported in Table 2 and Table 3, respectively highlighting the languages with human-annotated parallel data provided by the PAN shared task organizers, and those without.

Our model reaches its highest performance on German (J = 0.798, 1st) and its lowest on Amharic (J = 0.446, 8th) among the first batch of languages. The model also peaks on French (J = 0.802, 1st) and falls to its minimum on Hinglish (J = 0.511, 1st), a low-resource language. Overall, as we will also outline below in Paragraph 5.3, the lowest results by far are reported in all of the low-resource languages.

Table 2
Automatic evaluation results across different model variants on languages with available parallel data from human annotators. The results refer to the final submission phase. In parentheses: rank per language.

Model	Avg J (rank)	en (rank)	es (rank)	de (rank)	zh (rank)	ar (rank)	hi (rank)	uk (rank)	ru (rank)	am (rank)
Our submission	0.685 (1)	0.734 (3)	0.686 (11)	0.798 (1)	0.618 (1)	0.718 (3)	0.619 (7)	0.799 (2)	0.749 (6)	0.446 (10)
baseline_mt0	0.675 (5)	0.727 (6)	0.696 (10)	0.757 (6)	0.543 (8)	0.715 (4)	0.627 (5)	0.770 (6)	0.754 (2)	0.491 (1)
baseline_delete	0.536 (26)	0.473 (29)	0.603 (26)	0.586 (25)	0.516 (15)	0.611 (16)	0.480 (25)	0.581 (26)	0.514 (28)	0.461 (5)
baseline_gpt4	0.637 (12)	0.708 (13)	0.708 (5)	0.728 (12)	0.513 (16)	0.603 (17)	0.605 (10)	0.747 (11)	0.706 (12)	0.412 (16)
baseline_backtranslation	0.481 (28)	0.684 (18)	0.528 (29)	0.513 (29)	0.290 (29)	0.438 (28)	0.419 (28)	0.498 (28)	0.696 (13)	0.265 (27)

Table 3
Automatic evaluation results across different model variants on languages with no available parallel data from human annotators. The results refer to the final submission phase. In parentheses: rank per language.

Model	Avg J (rank)	it (rank)	ja (rank)	he (rank)	fr (rank)	tt (rank)	hin (rank)
Our submission	0.643 (1)	0.784 (1)	0.674 (1)	0.531 (1)	0.802 (1)	0.556 (5)	0.511 (1)
baseline_mt0	0.572 (12)	0.746 (8)	0.582 (13)	0.415 (23)	0.760 (9)	0.580 (4)	0.351 (18)
baseline_delete	0.510 (21)	0.668 (21)	0.441 (26)	0.436 (18)	0.518 (29)	0.573 (7)	0.425 (9)
baseline_gpt4	0.579 (9)	0.742 (10)	0.637 (7)	0.513 (3)	0.780 (6)	0.468 (17)	0.333 (19)
baseline_backtranslation	0.342 (20)	0.462 (29)	0.241 (31)	0.339 (28)	0.626 (26)	0.254 (30)	0.133 (30)

Interestingly, the baseline mT0 model [2] already achieves very competitive joint scores across most languages, ranking within the top 5 for 9 of the 17 tested languages and topping the leader-board on Amharic (0.491, 1st) despite zero fine-tuning. Its overall average J_P of 0.675 places it only four places below our final system, underscoring the effectiveness of large, multilingual foundational models out of the box. GPT-4 also performs respectably (Avg J_P 0.637, 12th), but falls behind MT0 on high-resource and cross-lingual pairs. These results highlight that simple, prompt-based baselines like MT0 remain strong contenders in detoxification tasks. For complete scores see Appendix C.

Moreover, Figure 1 illustrates the progression of the joint score J through our experiments and model development, through successive Gemma variants. The data refer to the Dev set simply for availability reasons, however, all data are perfectly comparable to our Test set as well. Starting from the 4B baseline (0.562 / 0.514), we see that with Few-shot prompting J rises by +0.022 on parallel data and falls by −0.015 on no-parallel data (to 0.584 / 0.499). By scaling up to the 12B variant and using few-shot approach, we reach a further gain of +0.081 (parallel) and +0.071 (no-parallel), amounting 0.643 / 0.585. By implementing Chain-of-Thought, J climbs to 0.650 / 0.592 (Δ +0.088 / +0.078 over baseline). During the two following phases (corresponding to the data augmentation phases) we register more modest improvements, ending at 0.685 / 0.607 and 0.692 / 0.642 respectively on the Dev set.

Figure 2 reports the evolution of our joint score J on the four lowest-resource languages (Amharic, Hebrew, Tatar and Hinglish) through each of the three fine-tuning phases (plus the zero-shot baseline, Phase 0). Amharic dips under few-shot CoT before modest recovery (Phase 0→3: 0.47→0.45), Hebrew gains steadily (0.48→0.53), and Tatar and Hinglish exhibit the largest rises, particularly after adding CoT and synthetic data (Tatar 0.43→0.56; Hinglish 0.40→0.51).

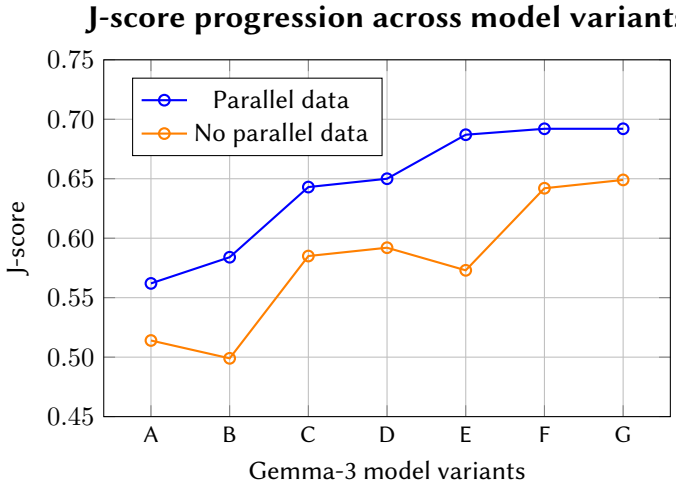

Figure 1: Model performance across phases: A = Gemma 4B zero-shot, B = Gemma 4B few-shot, C = Gemma 12B few-shot, D= Gemma 12B few-shot (additional phase), E = Gemma 12B CoT and few-shot, F = Gemma 12B CoT and few-shot (additional phase), G = Gemma 12B CoT and few-shot (second additional phase).

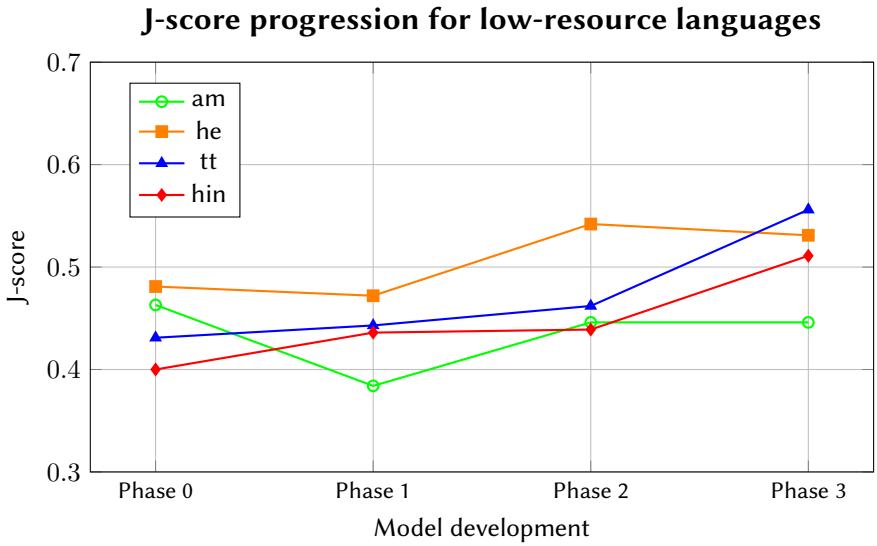

Figure 2: J-scores across four phases of our development of Gemma-3 12B for detoxification, specifically of low-resource and worst-achieving languages: Amharic (am), Hebrew (he), Tatar (tt), Hinglish (hin).

5.1. Additional results from LLM-as-a-Judge

After the end of the competition, the shared-task committee re-evaluated every submission with an LLM-as-a-Judge protocol intended to offer advanced means of analysis to the proposed models. The committee fine-tuned Llama-3.1-8B-Instruct [30] on the pairwise human annotations released for the same challenge in 2024 [4][6]. The fine-tuned model outputs a non-toxicity preference score that replaces the XLM-R probability p in equation (1) above (4.1), and a content-similarity score that replaces the LaBSE mixture in equation (2) above. Fluency continues to employ xCOMET as before. The joint score is therefore computed as:

$$J = X_{\text{comet\_fluency}}(\text{input}, \text{output}_{\text{gold}}, \text{output}_{\text{gen}}) \times \text{SIM}_{\text{LLM}}(\text{input}, \text{output}_{\text{gold}}, \text{output}_{\text{gen}}) \times \text{STA}_{\text{LLM}} \quad (4)$$

We provide results in 4 and 5: there, we include the committee's Golden Annotation performance,

[6] https://github.com/textdetox/textdetox_clef_2024/tree/main/human_evaluation_results

as well as the top two teams which ranked first in the subgroups of languages with and without the available parallel data. We additionally include the best performing baseline in each subgroup. Unlike the original automatic evaluation where our model ranked first on several languages, the LLM-as-a-Judge protocol demotes our system to third in both subgroups. The average J score across all fifteen languages would place us in second with $J = 0.7669$: namely, significantly below the Golden Annotation score ($J = 0.8233$) but slightly in front of the "MetaDetox" team ($J = 0.7636$) and "ReText.Ai" team ($J = 0.7538$).

Table 4
Results with the LLM-as-a-Judge protocol on languages with available parallel data from human annotators.

Model	Avg J (rank)	en	es	de	zh	ar	hi	uk	ru	am
Golden annotation	0.820 (1)	0.846	0.783	0.930	0.716	0.838	0.888	0.807	0.828	0.742
"MetaDetox" team	0.812 (2)	0.893	0.823	0.919	0.813	0.826	0.785	0.791	0.829	0.626
Our model	0.798 (3)	0.871	0.797	0.919	0.796	0.814	0.762	0.785	0.827	0.614
"ReText.Ai" team	0.775 (4)	0.794	0.765	0.888	0.783	0.790	0.773	0.791	0.792	0.597
baseline_mt0	0.768 (8)	0.843	0.764	0.825	0.717	0.791	0.751	0.770	0.809	0.639

Table 5
Results with the LLM-as-a-Judge protocol on languages with no available parallel data from human annotators.

Model	Avg J (rank)	it	ja	he	fr	tt	hin
Golden annotation	0.828 (1)	0.893	0.904	0.783	0.724	0.780	0.887
"ReText.Ai" team	0.722 (2)	0.823	0.805	0.657	0.860	0.583	0.606
Our model	0.720 (3)	0.842	0.820	0.681	0.889	0.495	0.592
"MetaDetox" team	0.691 (5)	0.821	0.721	0.610	0.883	0.493	0.621
baseline_gpt4	0.662 (11)	0.790	0.779	0.578	0.865	0.438	0.524

The LLM acting "as judge" was trained on pairwise human preference data, and thus acts as a proxy for real-world acceptability Our model's performance drop seems to suggest that human-calibrated LLM judgments emphasize different dimensions of text style, and a future investigation of such differences will certainly yield useful insights. Overall, we believe that our model proved to be sufficiently robust as it maintained a competitive standing even under this additional protocol. At the same time, these results seem to reinforce the need for detoxification systems to be optimized beyond standard fluency or similarity metrics.

5.2. Qualitative sampling analysis

For each of the fifteen target languages, we randomly sampled and manually analyzed 150 input–output sentence pairs from the system's submissions. Our qualitative review focused on identifying common error patterns, understanding their linguistic or resource-related drivers, and illustrating representative examples. We organize our observations into three main trends.

(i) Compound toxicity detection failures. In low- and medium-resource languages, the model frequently fails to identify toxicity when multiple toxic elements co-occur in the input. We find that this phenomenon is most pronounced in the Hinglish subset (see Appendix D) occurring in 6.7% of output sentences, likely due to mixed-script and code-switching which are fairly frequent in the language corpus. We also observe similar failures in Spanish and Mandarin Chinese: regional or sociolinguistically connotated profanities are often missed, leading to incomplete detoxification[7]. We attribute these errors to the extensive dialectal diversity in these two languages, which, with approximately one billion and half a billion native speakers respectively, encompass a broad vocabulary of localized curse terms. Representative examples are provided in Table 6, where the model repeatedly failed to detect the

[7]Of this type of errors, our qualitative analysis only detected one example for Spanish and one for Mandarin Chinese.

Table 6
Examples of missed toxic elements in Spanish.

Language	Original	Detoxified
Spanish	Udec culia se demoran ene en sus trámites culiaos, vale pico!	Udec se demoran mucho en sus trámites, vale pico!
Spanish	Mi perra vale pico, juro que me había levantado a jugar y tomó su peluche jajaajja Küyén de mrda	Mi novia vale pico, juro que me había levantado a jugar y tomó su peluche jajaajja Que asco.

Table 7
Examples of missed toxic elements in Spanish.

Language	Original	Detoxified
French	et en défendant comme des salopes, j'espère bien que vous gagnez !	et en défendant comme des idiots, j'espère bien que vous gagnez !
Italian	Ora ve lo dico veramente: VAFFANCULO A TUTTI VOI. Siete solo degli sfigati: bannatemi pure ma non me ne fotte un cazzo, coglioni!!!!	Ora ve lo dico veramente: Sono arrabbiato con tutti voi. Siete solo dei perdenti: bannatemi pure ma non me ne importa niente, idioti!!!!

Table 8
ANOVA: summary and effect sizes on Test set

Method	η^2	F-statistic	df (between, within)	p-value
Genetic affiliation	0.592	8.717	(2, 12)	0.00459
Typology	0.654	3.403	(5, 9)	0.05303
Geographical proximity	0.441	1.772	(4, 9)	0.21845
Language-resource tier	0.667	12.040	(2, 12)	0.00135

expression "vale pico", a toxic word for "worthless" that is reported to be in use in Southern American language communities, especially in Argentina.

(ii) Nonsensical paraphrase generation in low-resource languages. Among the lowest-resource languages, Amharic exhibits a disproportionately high rate of nonsensical outputs despite the availability of high-quality human-parallel data in the test corpus. Roughly 30% of sampled Amharic pairs yielded ungrammatical or semantically empty sentences.

(iii) Ambiguity in toxicity definition exacerbates errors. The absence of clear-cut boundaries for "toxicity" amplifies modeling challenges, particularly for languages where the input data is dominated by political discourse. Hebrew is the clearest example: coarse annotations in the test set combine explicit insults, political disapproval, and partisan hate speech. In this context, identifying only explicit toxic terms is difficult, and models often either over-detoxify (removing non-toxic political criticism) or under-detoxify (leaving implicit hostility intact). Moreover, the fine boundaries of toxicity lead the model to substitute words that straddle polite and impolite registers (e.g., terms like "idiot", which are common across Indo-European languages). Such substitutions result in outputs that technically remove overt insults but replace them with milder yet still demeaning vocabulary; see Table 7.

5.3. ANOVA analysis

We performed one-way ANOVA on mean detoxification scores across 15 languages using four grouping schemes: genetic affiliation, typology, geographical proximity, and language-resource status. Our methodology and criteria are outlined in Appendix B. Singleton-language clusters were aggregated into a unified "Other" category or omitted to ensure each factor level contained at least two members. Table 8 summarizes the Test-set results.

The ANOVA shows that language-resource status is the strongest predictor of detox performance, explaining two-thirds of the variance ($\eta^2 = 0.667$, $F(2, 12) = 12.04$, $p = 0.00135$). Genetic affiliation

also has a statistically significant effect, accounting for nearly 60% of variance ($\eta^2 = 0.592$, $F(2, 12) = 8.72$, $p = 0.00459$), indicating that shared language family membership correlates with similar detox behavior. Morphological typology exhibits a large effect size ($\eta^2 = 0.654$) but narrowly fails to reach the 5% significance threshold ($F(5, 9) = 3.40$, $p = 0.053$), suggesting a strong trend that may warrant further study[8]. By contrast, geographical proximity shows a weaker, non-significant effect ($\eta^2 = 0.441$, $F(4, 9) = 1.77$, $p = 0.218$), implying that regional grouping alone does not reliably predict detox performance once other factors are accounted for.

6. Limitations

Although fine-tuning and curated data availability are essential, our system still largely depends on the pretrained model's base capabilities; in the automatic evaluation phase, the zero-shot mT0 baseline scores 0.675 on the joint metric compared to our system's 0.685, showing only modest improvements from task-specific adaptation.

While prompts were manually translated and validated, ensuring perfect cross-linguistic consistency remains challenging. As pointed by one of our anonymous reviewers, subtle semantic shifts in phrasing may skew model behavior across languages. Additionally, the LaBSE-based retrieval of semantically similar sentences (3.1.3) can be unreliable for very low-resource languages and can likely compromise the quality of the few-shot examples.

An anonymous reviewer also rightly noted that our vast reliance on synthetic data (3.1) may introduce semantic bias and encourage stylistic overfitting. Synthetic text clearly lack the diversity of human-authored examples and likely limits the robustness and generalizability of the fine-tuned system.

Furthermore, on bias specifically, we acknowledge that like most NLP systems of the same type, our detoxification model does not perform equally across social-media posts from diverse cultural, religious, or socioeconomic domains. Performance disparities arise from several sources of bias. First, our system builds on an opaque foundation model (Gemma3), whose pre-training corpus is neither public nor documented. Any skew toward well-resourced language varieties is locked into the representations we start from, and there is no fully reliable way to audit, retrace, or rebalance those original samples. Unlike tasks where data can be recollected, this selection bias is structural and irreversible: we can only mitigate downstream effects but cannot avoid bias at its source.

In fact, our qualitative review already identified missed dialectal expressions; for example, Southern-American Spanish slang "vale pico" was systematically undetected. Overall, we believe that robust, consistent text detoxification across cultural, religious, socioeconomic, and other demographics requires bias-aware adaptations at both the model pre-training, post-training, and evaluation stages.

7. Future directions

The well-defined evaluation metrics of detoxification tasks make them ideal candidates for preference-based optimization techniques such as GRPO and DPO, which directly optimize non-toxicity while preserving meaning. Future research could explore integrating such approaches with multilingual LLMs to fine-tune systems more effectively in line with human preferences. Furthermore, to address the growing challenge of creative obfuscation (e.g., masked profanities like "f@ck," "@ss" as well as emoji-based strategies) and emergent slang, especially in low-resource settings, detoxification systems should incorporate dynamic lexicon updating and character-substitution detection.

We also see promise in a multi-stage reasoning pipeline of the following type: binary toxicity classification → toxicity-type prediction → targeted detoxification → quality verification. It will be particularly interesting to explore such a pipeline and the emerging reasoning language models (RLMs) as they become more accessible to the research community. Additionally, domain-informed data

[8]Modern historical linguistics has largely shown how genetic affiliation does not yield typological alignment since, diachronically, languages often fail to retain their core grammatical and structural features.

augmentation, which generates synthetic training examples that closely mirror test-time phenomena, might offer a practical path to more robust and generalizable fine-tuning corpora.

Lastly, our ANOVA results suggest that both morphological typology (η = 0.654) and genetic affiliation (η = 0.592) together explain the majority of performance variance in detoxification tasks. However, broader language sampling and mixed-effects modeling are necessary to disentangle their actual contributions more clearly.

Acknowledgments

The authors would like to thank the Faculty members and all fellow students of the Master of Science program in "IT & Cognition" at the University of Copenhagen, for their feedback and useful suggestions. We are especially grateful to Professor Patrizia Paggio and Professor Manex Agirrezabal for introducing us to the field of automatic text detoxification.

Declaration on Generative AI

During the preparation of this work, the authors used Generative AI assistance tools for: grammar and spelling review; synthetic data generation (as outlined above in the Methods section). The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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.

[2] D. Dementieva, V. Protasov, N. Babakov, N. Rizwan, I. Alimova, C. Brune, V. Konovalov, A. Muti, C. Liebeskind, M. Litvak, D. Nozza, S. Shah Khan, S. Takeshita, N. Vanetik, A. A. Ayele, F. Schneider, X. Wang, S. M. Yimam, A. Elnagar, A. Mukherjee, A. Panchenko, Overview of the multilingual text detoxification task at pan 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[3] C.-M. Pascale, The weaponization of language: Discourses of rising right-wing authoritarianism, Current Sociology 67 (2019). doi:10.1177/0011392119869963, (Original work published 2019).

[4] 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: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, Grenoble, France, 2024, pp. 2432–2461.

[5] Toxic comment classification challenge dataset, Kaggle competition dataset, 2018. URL: https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge.

[6] C. Nogueira dos Santos, I. Melnyk, I. Padhi, Fighting offensive language on social media with unsupervised text style transfer, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 189–194. URL: https://aclanthology.org/P18-2031/. doi:10.18653/v1/P18-2031.

[7] D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, A. Panchenko, Text detoxification using large pre-trained neural models, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7979–7996. URL: https://aclanthology.org/2021.emnlp-main.629/. doi:10.18653/v1/2021.emnlp-main.629.

[8] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, A. Panchenko, Paradetox: Detoxification with parallel data, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), 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.

[9] A. Pesaranghader, N. Verma, M. Bharadwaj, Gpt-detox: An in-context learning-based paraphraser for text detoxification, 2024. URL: https://arxiv.org/abs/2404.03052. doi:10.1109/ICMLA58977.2023.00230. arXiv:2404.03052.

[10] 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/. doi:10.18653/v1/2024.naacl-short.12.

[11] S. Teo, The co-star framework for prompt engineering, Blog post, *Towards Data Science*, 2023. URL: https://towardsdatascience.com/how-i-won-singapores-gpt-4-prompt-engineering-competition-34c195a93d41.

[12] 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. Di 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/.

[13] G. Floto, M. M. Abdollah Pour, P. Farinneya, Z. Tang, A. Pesaranghader, M. Bharadwaj, S. Sanner, Diffudetox: A mixed diffusion model for text detoxification, 2023. URL: https://arxiv.org/abs/2306.08505. arXiv:2306.08505.

[14] R. Shah, Rlm–hinglish translator, https://huggingface.co/rudrashah/RLM-hinglish-translator, 2022. Accessed: 2025-05-28.

[15] NLLB Team, No Language Left Behind: Building a scalable and inclusive multilingual translation system, arXiv preprint arXiv:2207.04672 (2022).

[16] A. D. Dang, contributors, Paradetox–mt6: Synthetic multilingual detoxification corpus, https://huggingface.co/datasets/anhdtd/augument_data, 2025. Version 1.0, accessed 28 May 2025.

[17] TextDetox Team, Multilingual toxicity dataset, https://huggingface.co/datasets/textdetox/multilingual_toxicity_dataset, 2024. Version 1.0, accessed 28 May 2025.

[18] D. Moskovskiy, N. Sushko, S. Pletenev, E. Tutubalina, A. Panchenko, Synthdetoxm: Modern llms are few-shot parallel detoxification data annotators, 2025. URL: https://arxiv.org/abs/2502.06394. arXiv:2502.06394.

[19] A. Zeira, J. Montal, Multilingual toxic lexicon, Hugging Face dataset, 2024. URL: https://huggingface.co/datasets/textdetox/multilingual_toxic_lexicon/, accessed: 2025-05-28.

[20] NLLB Team, facebook/nllb-200-distilled-600m: A distilled subset of nllb-200, Hugging Face model, 2023. URL: https://huggingface.co/facebook/nllb-200-distilled-600M, accessed: 2025-05-28.

[21] S-NLP Research, s-nlp/bart-base-detox: Bart-base model fine-tuned for english detoxification, Hugging Face model, 2023. URL: https://huggingface.co/s-nlp/bart-base-detox, accessed: 2025-05-28.

[22] M. Labonne, mlabonne/llama-3.1-70b-instruct-lorablated: Llama-3.1 70b instruction-

tuned variant, Hugging Face model, 2024. URL: https://huggingface.co/mlabonne/Llama-3.1-70B-Instruct-lorablated, accessed: 2025-05-28.

[23] OpenAI, Gpt-4-0613, gpt-4o-2024-08-06, and o3-mini-2025-01-31: proprietary large language models, OpenAI API, 2025. Accessed: 2025-05-28.

[24] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai, A. Tsitsulin, R. Busa-Fekete, A. Feng, N. Sachdeva, B. Coleman, Y. Gao, B. Mustafa, I. Barr, E. Parisotto, D. Tian, M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agarwal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen, A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Feng, A. Kolesnikov, A. Bendebury, A. Abdagic, A. Vadi, A. György, A. S. Pinto, A. Das, A. Bapna, A. Miech, A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini, C. Chen, C. L. Lan, C. A. Choquette-Choo, C. Carey, C. Brick, D. Deutsch, D. Eisenbud, D. Cattle, D. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle, E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi, H. Klimczak-Plucińska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne, I. Szpektor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. Wieting, J. Lai, J. Orbay, J. Fernandez, J. Newlan, J. yeong Ji, J. Singh, K. Black, K. Yu, K. Hui, K. Vodrahalli, K. Greff, L. Qiu, M. Valentine, M. Coelho, M. Ritter, M. Hoffman, M. Watson, M. Chaturvedi, M. Moynihan, M. Ma, N. Babar, N. Noy, N. Byrd, N. Roy, N. Momchev, N. Chauhan, N. Sachdeva, O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa, P. Xu, P. Stanczyk, P. Tafti, R. Shivanna, R. Wu, R. Pan, R. Rokni, R. Willoughby, R. Vallu, R. Mullins, S. Jerome, S. Smoot, S. Girgin, S. Iqbal, S. Reddy, S. Sheth, S. Põder, S. Bhatnagar, S. R. Panyam, S. Eiger, S. Zhang, T. Liu, T. Yacovone, T. Liechty, U. Kalra, U. Evci, V. Misra, V. Roseberry, V. Feinberg, V. Kolesnikov, W. Han, W. Kwon, X. Chen, Y. Chow, Y. Zhu, Z. Wei, Z. Egyed, V. Cotruta, M. Giang, P. Kirk, A. Rao, K. Black, N. Babar, J. Lo, E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. Mirrokni, E. Senter, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, Y. Matias, D. Sculley, S. Petrov, N. Fiedel, N. Shazeer, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, J.-B. Alayrac, R. Anil, Dmitry, Lepikhin, S. Borgeaud, O. Bachem, A. Joulin, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, Gemma 3 technical report, 2025. URL: https://arxiv.org/abs/2503.19786. arXiv:2503.19786.

[25] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.

[26] H. Face, xlmr-large-toxicity-classifier-v2, https://huggingface.co/textdetox/xlmr-large-toxicity-classifier-v2, 2023. Accessed: 2025-05-28.

[27] H. Face, Labse, https://huggingface.co/sentence-transformers/LaBSE, 2023. Accessed: 2025-05-28.

[28] H. Face, Xcomet-lite, https://huggingface.co/myyycroft/XCOMET-lite, 2023. Accessed: 2025-05-28.

[29] 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: Proceedings of the Tenth International Conference on Learning Representations (ICLR), 2022. ArXiv:2106.09685.

[30] Meta AI, Llama-3.1-8B-Instruct, https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, 2024. Model version 3.1, accessed 6 Jul 2025.

[31] C. A. Ciancaglini, How to prove genetic relationships among languages: The cases of japanese and korean, Rivista Degli Studi Orientali (2008). URL: https://doi.org/10.1400/143140. doi:10.1400/143140.

[32] E. A. Moravcsik, Introducing Language Typology, Cambridge Introductions to Language and Linguistics, illustrated ed., Cambridge University Press, Cambridge, 2013. doi:10.1017/CBO9780511978876.

[33] G. Son, D. Yoon, J. Suk, J. Aula-Blasco, M. Aslan, V. T. Kim, S. B. Islam, J. Prats-Cristià, L. Tormo-Bañuelos, S. Kim, Mm-eval: A multilingual meta-evaluation benchmark for llm-as-a-judge and reward models, 2025. URL: https://arxiv.org/abs/2410.17578. arXiv:2410.17578.

[34] T. Zhong, Z. Yang, Z. Liu, R. Zhang, Y. Liu, H. Sun, Y. Pan, Y. Li, Y. Zhou, H. Jiang, J. Chen, T. Liu, Opportunities and challenges of large language models for low-resource languages in humanities research, 2024. URL: https://arxiv.org/abs/2412.04497. arXiv:2412.04497.

A. Appendix: Data augmentation through machine translation

Table 9
Summary of the machine-translation models used for data augmentation.

Model	Architecture (params)	Training corpus	Tokenizer
RLM–Hinglish Translator	Gemma-2B causal LM; LoRA-fine-tuned	100 M Hinglish–English sentence pairs	64 k BPE
NLLB-200	24-layer encoder–decoder Transformer; 3.3 B params	4.4 B multilingual sentences (200 langs.)	256 k SentencePiece

B. Appendix: Grouping methodology for ANOVA analysis

B.1. Genetic affiliation

Eight Indo-European languages (English, Spanish, German, Hindi, Ukrainian, Russian, Italian, French) formed the largest cluster. The three Semitic tongues (Arabic, Amharic, Hebrew) comprised a second, genetically coherent group. Chinese stood alone as Sino-Tibetan, Tatar as Turkic. Japanese was also treated as unclassified since, unlike common misconceptions in modern historical linguistics, no definitive evidence has been found linking Japanese to any other linguistic family [31]. Hinglish, a contemporary English–Hindi pidgin, was non included in the Indo-European cluster. Therefore the latter four languages (Chinese, Tatar, Hinglish, Japanese) were combined into an "Other" group of unrelated members.

B.2. Typology

We grouped languages by their dominant word-order and morphological processes, with the latter features being prioritized [32]. The “fusional + SVO” cluster included the Western Romance languages (Spanish, Italian, French). Semitic root-and-pattern morphology defined the “fusional + templatic” group (Arabic, Amharic, Hebrew). The case-rich fusional-synthetic category (of "case-languages") comprised Ukrainian, Russian, and German. Agglutinative morphology linked Tatar and Japanese. We also tested an “isolating + English” cluster with English and Chinese, since English exhibits primarily analytic structures with some minor isolating features; Chinese itself formed a pure isolating group. Hinglish and Hindi were left unclustered here: Hinglish as a pidgin does not conform easily to traditional morphological typologies, and Hindi’s mixed agglutinative/isolating profile resisted assignment to a single category. Therefore Chinese, Hindi, and Hinglish were combined into an "Other" group of unrelated members.

B.3. Geographical proximity

We partitioned languages by broad region. Western Europe encompassed English, Spanish, German, Italian, and French. Eastern Europe and Eurasian Russia grouped Ukrainian, Russian, and Tatar. The Middle East paired Arabic and Hebrew, while North/East Africa was represented solely by Amharic (omitted in ANOVA). South Asia included Hindi and Hinglish, and East Asia combined Chinese and Japanese.

B.4. Language-resource tier

Languages were partitioned according to [33] [34]: eight high-resource languages (English, Chinese, Spanish, German, French, Russian, Italian, Japanese) formed the first cluster; a medium-resource group comprised Arabic, Hindi, Hebrew, and Ukrainian; and the low-resource tier included Amharic, Tatar, and Hinglish.

C. Appendix: Baselines results

Table 10
Complete automatic evaluation results for baseline submissions.

Model	AvgP	en	es	de	zh	ar	hi	uk	ru	am	AvgNP	it	ja	he	fr	tt	hin
mt0	0.675 (5)	0.727 (6)	0.696 (10)	0.757 (6)	0.543 (8)	0.715 (4)	0.627 (5)	0.770 (6)	0.754 (2)	0.491 (1)	0.572 (12)	0.746 (8)	0.582 (13)	0.415 (23)	0.760 (9)	0.580 (4)	0.351 (18)
gpt4	0.637 (12)	0.708 (13)	0.708 (5)	0.728 (12)	0.513 (16)	0.603 (17)	0.605 (10)	0.747 (11)	0.706 (12)	0.412 (16)	0.579 (9)	0.742 (10)	0.637 (7)	0.513 (3)	0.780 (6)	0.468 (17)	0.333 (19)
o3mini	0.562 (22)	0.688 (16)	0.660 (18)	0.607 (23)	0.439 (24)	0.498 (26)	0.549 (21)	0.685 (21)	0.638 (23)	0.291 (26)	0.484 (25)	0.605 (28)	0.490 (24)	0.475 (11)	0.725 (17)	0.360 (26)	0.251 (24)
gpt4o	0.560 (23)	0.615 (27)	0.656 (20)	0.572 (26)	0.391 (27)	0.529 (24)	0.547 (22)	0.706 (17)	0.646 (22)	0.379 (19)	0.535 (16)	0.677 (18)	0.567 (16)	0.451 (15)	0.709 (19)	0.443 (20)	0.362 (16)
delete	0.536 (26)	0.473 (29)	0.603 (26)	0.586 (25)	0.516 (15)	0.611 (16)	0.480 (25)	0.581 (26)	0.514 (28)	0.461 (5)	0.510 (21)	0.668 (21)	0.441 (26)	0.436 (18)	0.518 (29)	0.573 (7)	0.425 (9)
backtranslation	0.481 (28)	0.684 (18)	0.528 (29)	0.513 (29)	0.290 (29)	0.438 (28)	0.419 (28)	0.498 (28)	0.696 (13)	0.265 (27)	0.342 (30)	0.462 (29)	0.241 (31)	0.339 (28)	0.626 (26)	0.254 (30)	0.133 (30)
duplicate	0.475 (29)	0.353 (30)	0.566 (28)	0.572 (27)	0.477 (22)	0.564 (20)	0.417 (29)	0.442 (29)	0.424 (30)	0.461 (7)	0.482 (26)	0.653 (23)	0.440 (27)	0.425 (20)	0.447 (30)	0.510 (11)	0.419 (10)

D. Appendix: Toxicity retention in Hinglish

Table 11
Hinglish outputs that still contain toxicity when more than one toxic element appears in the input.

Language	Input (toxic)	Output (still toxic)
Hinglish	Haa bhn cho sb chutiye hai koi ni sunta	Haan, sab chutiye hai, koi ni sunta.
Hinglish	Yeh Budda ko kisne Glucose chadaya? BC ko valium 10 de do.	Yeh Budda ko kisne Glucose chadaya? BC ko valium 10 de do.
Hinglish	BakcHodi na failao ab randi ke aulad	BakcHodi na failao ab.
Hinglish	Mukul sudhar jaa warna bhosdike randi ki aulad sale teri bhen ko sare aam chod denge tu kuch nhi kar payga	Mukul sudhar jaa warna teri maa ko sare aam chod denge tu kuch nhi kar payga
Hinglish	Hmare desh k neta motherchood hai. Ye sale sirf ninda krte hai .	Hmare desh k neta behtar ho sakte hai. Ye sale sirf ninda krte hai .