TextDetox CLEF 2025/Multilingual Text Detoxification 2025 Jiaozipi: A Multilingual Text Detoxification Method Based on Large Language Model-Based Ensemble Learning

Notebook for the PAN Lab Lab at CLEF 2025

Xiaohui Liu[†], Yusheng Yi[*], Zhaotian Chen[†], Simin Xu, Zijun Ke, Xin Guo, Yubo Huang, Wenxuan Zhang, Jiayi Chen and Yong Han[*]

Foshan University, Foshan, China

Abstract

This paper proposes a solution for the multilingual text detoxification task at CLEF 2025. The task requires detoxification of explicit toxic texts across 15 languages while saving the main content as much as possible. To address the task, we propose a solution based on prompt engineering and ensemble of LLMs. As a first step, we extend the official dataset to construct a parallel text detoxification dataset and a toxic keywords list. We first employ the RISE prompting framework to generate initial system instructions. These instructions, combined with few-shot examples and user input, form structured prompts that guide multiple commercial large language models (DeepSeek, Qwen, Kimi) to produce detoxified outputs. Finally, the best results are selected via multi-dimensional evaluation considering semantic preservation, toxicity reduction, style consistency, and fluency. Our method is ranked 9th in automatic evaluation metrics.

Keywords

PAN 2025, multilingual text detoxification, large language model, RISE, Few-shot Learning

1. Introduction

With the rapid development of social media, toxic texts on online platforms have increased sharply, including racial discrimination remarks, personal attacks, hate speech, and other inappropriate content. To address this issue, text detoxification has been proposed as an intervention approach grounded in natural language generation. The advanced approach of the text detoxification primarily employs deep learning models to automatically detect toxic elements in text, such as insulting or discriminatory expressions.Then deep learning models transform them into neutral formulations that preserve the original semantic [1].

The task of multilingual text detoxification at CLEF 2025 [2, 3] aims at presenting a neutral version of a user message which preserves the original meaning. This task covers 15 languages, including high-resource languages such as English, Chinese, and Spanish, as well as low-resource or morphologically complex languages such as Amharic and Tatar.

The challenge of this task is implicit types toxiciy —like sarcasm, passive aggressiveness, or direct hate to some group where no neutral content can be found. Such implicit toxicity types are challenging to be detoxified so the intent will indeed become non-toxic.

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

[*]Corresponding authors: Yi is the first corresponding author and Han is the second corresponding author.

[†]These authors contributed equally.

1783226038@qq.com (X. Liu); yiys@fosu.edu.cn (Y. Yi); 1353663548z@gmail.com (Z. Chen); xuximi77@gmail.com (S. Xu); kzj3076418134@outlook.com (Z. Ke); gx13669565561@outlook.com (X. Guo); 1779920653@qq.com (Y. Huang); 3104881826@qq.com (W. Zhang); orangejy37@gmail.com (J. Chen); hanyong2005@fosu.edu.cn (Y. Han)

0009-0009-6571-9669 (X. Liu); 0009-0006-7098-3681 (Y. Yi); 0009-0008-3734-7442 (Z. Chen); 0009-0003-9829-5141 (S. Xu); 0009-0009-2300-2918 (Z. Ke); 0009-0006-0204-0024 (X. Guo); 0009-0006-2421-3766 (Y. Huang); 0009-0006-2534-7272 (W. Zhang); 0009-0007-6497-3687 (J. Chen); 0000-0002-9416-2398 (Y. Han)

2. Relate Work

Text detoxification tasks aim to convert toxic text into neutral expressions while preserving the original meaning. In 2024, Peng et al. proposed a method based on a few-shot learning and the CO-STAR framework, combined with chat models like Kimi for multilingual text detoxification. By generating few-shot learning contexts and structured prompts, this approach significantly improved the detoxification performance in high-resource languages like English and Chinese. [4]. In the same year, Řehulka and Šuppa explored retrieval-augmented generation (RAG) and dynamic prompt construction to enrich large language models (LLMs) with external knowledge, achieving competitive results in multilingual detoxification tasks [5]. However, for some low-resource languages such as Amharic, the lack of sufficient training data substantially limited performance. Therefore, they adopted a deletion approach, directly removing toxic keyword to ensure detoxification effectiveness. These approaches demonstrate notable progress in detoxification for high-resource languages, but their effectiveness remains constrained by limited multilingual training data. Effectively leveraging existing data to improve detoxification performance on low-resource languages remains a challenge.

LLMs, pretrained on massive corpora via self-supervised learning, acquire broad and emergent linguistic capabilities. However, achieving strong performance on specific downstream tasks often necessitates fine-tuning, which require substantial annotated data and computational resources. In contrast, prompt engineering enables the activation of latent model capabilities by designing effective prompt instructions. The prompt instructions can improve the relevance, coherence, and accuracy of model outputs [6].

Prompt engineering is a systematic approach to designing, writing, and optimizing input prompts for LLMs to guide them in producing expected output. To enhance the effectiveness of prompt engineering, various prompt frameworks have been proposed. For example, chain-of-thought (CoT) [7] and few-shot prompting [8] improve the interpretability and adaptability of LLMs in logical reasoning tasks and low-data scenarios by guiding models to break down complex problems or provide example references.

Although prompt engineering has enhanced the ability of LLMs to perform text detoxification tasks, single models still face challenges such as output instability and residual toxicity. Ensemble learning is a method that integrates the predictions of multiple base models to improve the robustness, accuracy, and generalization ability of the system [9]. However, existing ensemble methods are often static, relying on simple strategies such as majority voting or average scoring, which limits their flexibility and effectiveness in complex generation tasks like text detoxification.

3. Datasets

In this task, we need to detoxify 15 languages. However, the provided parallel text detoxification dataset[1] [10, 11, 12] covers only 9 languages. Therefore, we used Yuanbao AI[2] to translate the English portion of the parallel text detoxification dataset into Italian, French, Hebrew, Hinglish, Japanese, and Tatar, with 100 translations for each language. The process we performed is shown in table 1.

Table 1
Multilingual Examples from a Toxic-to-Neutral Sentence Alignment Corpus

toxic_sentence	neutral_sentence	lang
Are you there d*ck, the war is back on!	Are you there? The war is back on.	en
Sei lì, figlio di p*ttana, la guerra è ricominciata!	Sei lì? La guerra è ricominciata.	it
Tum waha ho kya, ch*tiye, jung phir se shuru ho gayi!	Tum waha ho? Jung phir se shuru ho gayi.	hin
T'es là, c*nnard, la guerre a recommencé !	T'es là ? La guerre a recommencé.	fr

[1] https://huggingface.co/datasets/textdetox/multilingual_paradetox
[2] https://yuanbao.tencent.com/

Although the coverage ability of current mainstream commercial LLMs on parallel text detoxification dataset has significantly improved, they still fail to recognize toxic keywords with cultural dependence, semantic ambiguity or distorted expressions.

To enhance the ability to recognize fine-grained toxic texts, we attempt to extract toxic words using the Toxic Keywords[3] [13, 14] provided in the task introduction. But it is insufficient to support the replacement of toxic text, because there are fewer words in it for example, Amharic toxic keywords have only 245 records. So we extract the negative words in Toxic Span[4] [15] and merged them with Toxic Keywords. The process that we extract negative words is shown in table 2.

Table 2
Toxic Keywords Extraction

Sentence	Negative Connotations
all you trump c*owns are seriously m*ssed up.	c*owns,m*ssed
allowing whole colonies of such r*bbish to arise should be p*nishable by f*ring the officials.	r*bbish, p*nishable, f*ring
almost as f*cked up as the cia funding and arming bin laden.	f*cked up
amy , your ignorance is showing again.	i*norance
and start sending cunts home.	c*nts

Note:Negative Connotations are what we need to extract and merge with Toxic Keywords .

Table 3
Summary of toxic keywords

lang	am	ar	de	en	es	hi	ru	uk	zh	it	ja	tt	fr	he	hin
quantity	1823	1322	1771	3884	1854	682	10000	7710	4606	815	328	10000	1287	731	209

As a first step, we generate a parallel text detoxification dataset and a toxic keywords list from the official dataset.

Ultimately, we obtained the datasets and toxic keywords lists of 15 languages, as follows:

- The extended datasets: There are 100 samples each for Italian(it), French(fr), Hebrew(he), Hinglish(hin), Japanese(ja), and Tatar(tt), and 400 samples each for English(en), Spanish(es), German(de), Chinese(zh), Arabic()ar, Hindi(hi), Ukrainian(uk), Russian(ru), and Amharicen(am). These samples will be provided as examples to the large model for the optimization of the model's output.
- Toxic Keywords List: The summary of each language entry is in Table 3. These toxic keywords will be replaced with * in the toxic sentence. The replaced sentence is called *the toxic_voc_replaced* result below.

4. Method

Our method consists of three main steps: 1) constructing prompt using the RISE framework, 2) inputting toxic sentences into three LLMs (Kimi[5], DeepSeek[6], and Qwen[7]) to generate detoxification results, 3) putting the detoxification results of the large models and *the toxic_voc_replaced* results into Qwen for quality evaluation and finally return the best result as the output.

[3]https://huggingface.co/datasets/textdetox/multilingual_toxic_lexicon
[4]https://huggingface.co/datasets/textdetox/multilingual_toxic_spans
[5]https://www.kimi.com
[6]https://chat.deepseek.com
[7]https://www.tongyi.com/

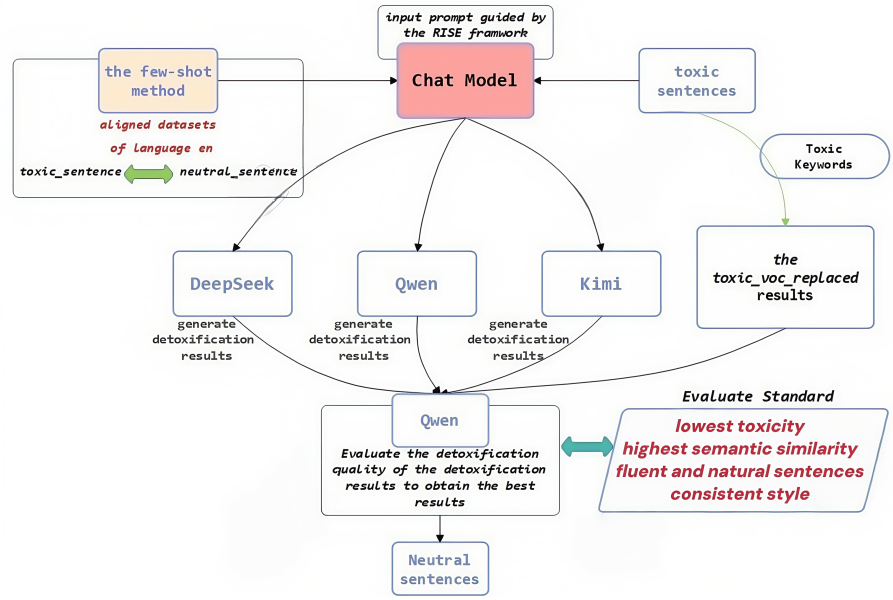

Figure 1: architecture of the detoxification model

4.1. Constructing input texts

1. Input prompt guided by The RISE framework

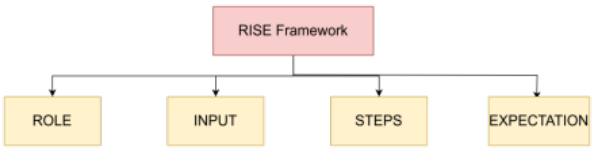

Figure 2: RISE framework

#Role#:
Assume you are an expert in language processing.

#Inputs#:
I have received a batch of toxic sentences (sentences containing harmful text) <|toxic_sentence|>
and their detoxified neutral counterparts (sentences with harmful text removed) <|neutral_sentence|>
in the context object.

#Steps#:
You can detoxify the sentences by removing harmful keywords or directly optimizing the sentences to convert them into neutral versions.

#Expectation of the result#:
Return the data in JSON format as follows:{"toxic_sentence": "", "neutral_sentence": "", "lang": "" }
Where:
lang is the language type: en, ru, uk, de, es, am, zh, ar, hi, it, fr, he, hin, tt, ja.
neutral_sentence is the detoxified neutral sentence, which should retain the original content while ensuring semantic similarity (measured by cosine similarity between LaBSE embeddings) and maintaining good fluency.
toxic_sentence is the original toxic sentence before detoxification.

Figure 3: display of detoxification prompt

Practical prompt construction is essential for eliciting optimal responses from LLMs. The RISE framework serves as our structured template for prompt design, as illustrated in Figure 2. Its

operational methodology for this task is delineated as follows: **Role (R):**The model is required to function as a domain expert in linguistic processing, specifically tasked with text detoxification. **Input (I):**The source material consists of toxic text along with supplementary contextual data, utilized for model training and refinement. **Steps (S):** A systematic approach—comprising keyword elimination and syntactic optimization—is employed to ensure precision and operational feasibility. **Expectation (E):**The output must preserve the original meaning while achieving semantic equivalence, linguistic fluency, and formal coherence. The response shall include a JSON output format like [*toxic_sentence*: "", *neutral_sentence*: "", lang: ""]

2. Generate few-shot learning context
   This section shows how we generate the contents of a few-shot learning context.

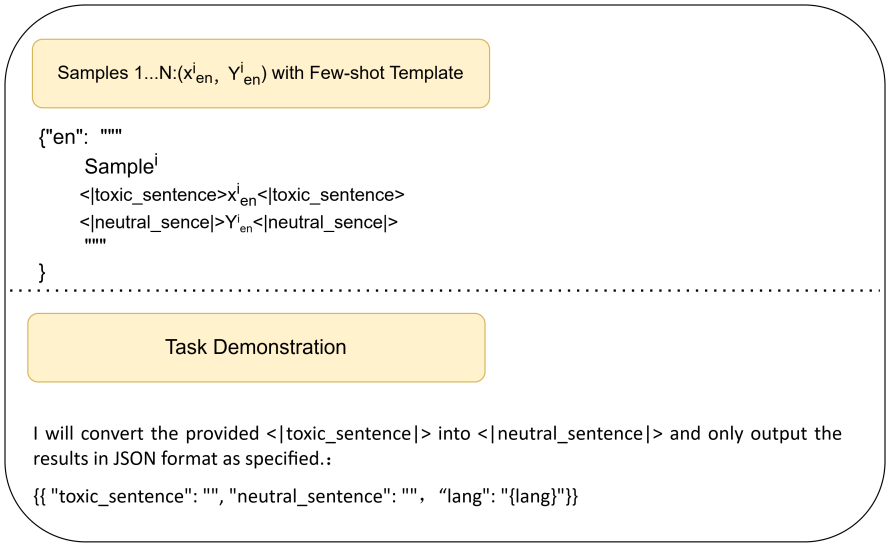

Figure 4: the generation of the few-shot learning

Task Demonstration: to assist LLMs in accurately understanding the task requirements, we provide a brief description of the task.
Few-shot learning content: to help the model understand the neutral version of toxic text, we provide few-shot learning content. This content contains toxic sentences and their corresponding neutral sentence pairs in parallel text detoxification dataset of the target language. Figure 4 shows an example of English (en), and the processing methods for other languages are the same. The parallel text detoxification dataset of each language is stored in dictionary form, making it easier to call up small sample learning content in the corresponding language later.

3. Input toxic sentences

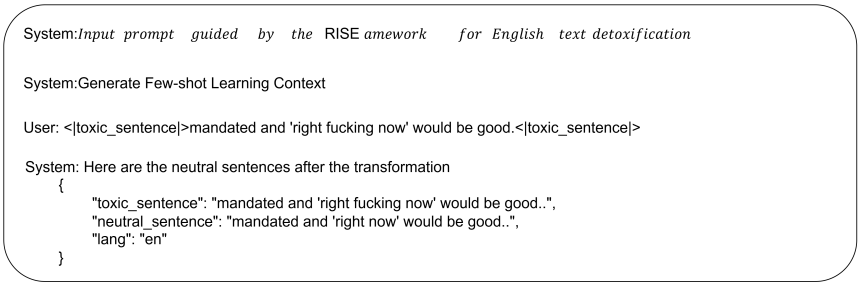

Figure 5: example of the detoxification process for large models

As Figure 5 shows, we insert a toxic sentence into template < |*toxic_sentence*| >< |*toxic_sentence*| >, and send it to the large language model. With the help of few-shot learning and prompts based

on the RISE framework, the large language model will return formatted neutral sentences, Figure 5 demonstrates the real detoxification process.

4.2. Evaluation:

In this section, we introduce how to use the Qwen model as an evaluator to evaluate and select optimal detoxification results.

1. Input Prompt of evaluation

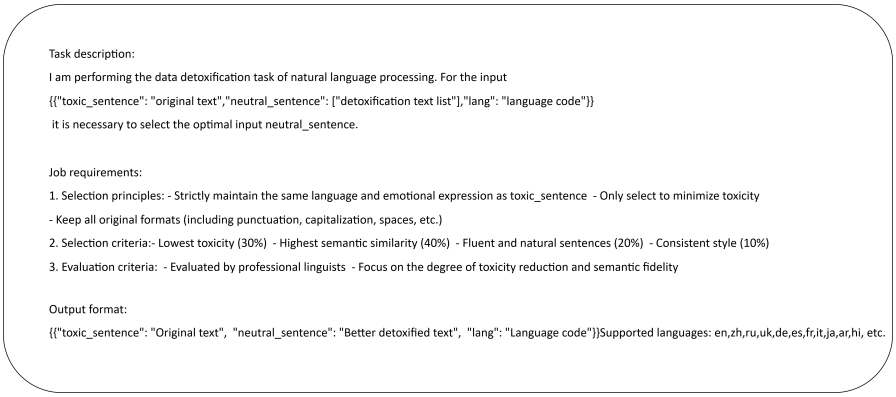

Figure 6: the display of evaluation prompt

This prompt involves selecting the optimal detoxified output from a list of candidate texts for a large language model. Our selection criteria and weightings are as follows: lowest toxicity score(weight: 0.3); highest semantic similarity to the original text (weight: 0.4); fluency and naturalness of the generated sentences(weight: 0.2); consistency in style(weight: 0.1). And we required a JSON output format like [*toxic_sentence*: "", *neutral_sentence*: "", lang: ""]

2. The evaluation process of large models

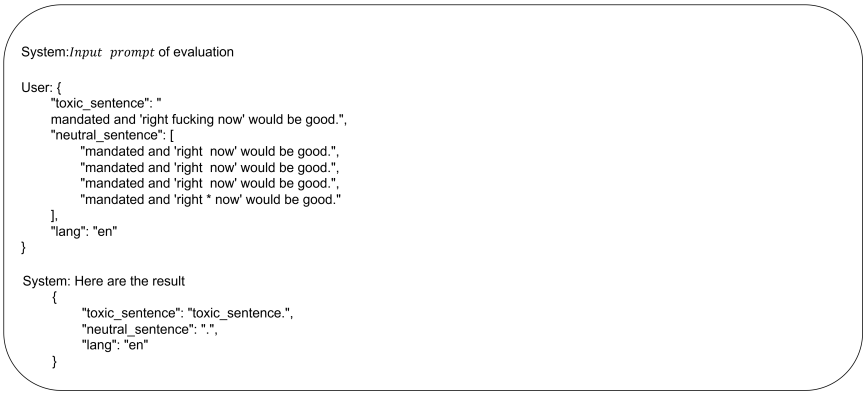

Figure 7: example of large model evaluation

We inserted the toxic sentence, the list of neutral sentences and corresponding language into template like Figure 7, and send it to the Qwen model. With the help of prompt, the large language model returns formatted neutral sentences. Figure 7 demonstrates the real valuation process.

5. Experiment

5.1. Settings

For all 15 languages, we repeat the following steps:

1. **Input of the few-shot learning context :** Construct the few-shot learning context using datasets and input it into the model. For different languages, replace the context sample information and language identifiers accordingly.
2. **Input prompts guided by the RISE framework :** Input the prompt words into the large model to guide it to generate the correct output.
3. **Input of toxic sentence :** Embed the toxic text between *<toxic sentence>* and *<toxic sentence>* of the framework (as shown in Figure 5), and then input it into the large language model.
4. **Evaluate:** The result of DeepSeek, Qwen and Kimi, and *the toxic_voc_replaced* results were input into the Qwen model for evaluation (as shown in Figure 7). Finally, best result as the output was returned.

5.2. Result

We applied our method to conduct systematic comparative experiments based on official datasets. In the comparative experiment of the prompt framework, in order to control the variables, the experiment first fixed the single model benchmark, and used DeepSeek for the large language model. the prompt framework combined with word replacement strategy was uniformly applied. We focus on the performance differences between the COSTAR framework and the RISE framework. Experimental data show that the RISE framework shows significant advantages in the core indicators, with an AvgP value of 0.636 and an AvgNP value of 0.565, compared with the corresponding index values of the COSTAR framework of 0.623 and 0.553 respectively (see Table 4 for details). Based on this empirical result, we decided to use the RISE framework as the prompt framework of the large model in the follow-up experiments to ensure the best detoxification effect of the experiment.

Table 4
Results of the prompt framework comparison experiment

Framework	AvgP	en	es	de	zh	ar	hi	uk	ru	am	AvgNP	it	ja	he	fr	tt	hin
COSTAR	0.623	0.677	0.669	0.714	0.466	0.658	0.611	0.753	0.694	0.369	0.553	0.677	0.523	0.462	0.721	0.474	0.460
RISE	0.636	0.680	0.679	0.717	0.512	0.679	0.612	0.746	0.713	0.388	0.565	0.695	0.577	0.484	0.721	0.455	0.456

Based on the RISE framework, we systematically compared the performance of four different text detoxification methods. As shown in Table 5, we evaluated the following methods in turn:

1. **Single large model + Prompt + Word replacement:** Using a single large model combined with prompt framework, targeted vocabulary replacement is performed on the parts with poor results to maintain the basic semantic structure after context processing;
2. **Single large model + Prompt + Back-translation:** Using a single large model combined with prompt framework, through cross-language conversion and secondary detoxification of the preliminary results to improve the effect of multilingual detoxification;
3. **Single large model + Prompt + Translation Detoxification:** Using a single large model combined with a prompt framework, first perform language conversion for the weak language of the model, then uniformly use the large model for detoxification, and finally translate the specific language back to the original language type;
4. **Multiple large models + Prompt + Word replacement:** Integrate the detoxification results output by multiple large models, and select the optimal detoxification text results in combination with word replacement;

As Table 5 shows that in the detoxification scheme using a single model, Strategy 1 which combines the prompt framework and word replacement strategy, exhibits the best decontamination effect. Compared to the other two methods, this method demonstrates significant advantages in six languages: German (de), Arabic (ar), Ukrainian (uk), Russian (ru), Tatar (tt), and Hinglish (hin), with both AvgP (0.636) and AvgNP (0.572) metrics outperforming those of other single model methods. In the follow-up, we compared it with the multi-model integrated detoxification method (Strategy 4) in comparative experiment. Further comparative experiment show that the multi-model ensemble detoxification method

achieves a breakthrough in the detoxification effect. Not only did the detoxification effect of French (fr) jump to 0.801, but it also surpassed the detoxification performance of a single model in all test languages except Amharic (am). This multilingual text detoxification method achieved the best experimental results so far, increasing the AvgP to 0.656 and the AvgNP to 0.607.

Table 5
Result of Comparative Experiments

Strategy	AvgP	en	es	de	zh	ar	hi	uk	ru	am	AvgNP	it	ja	he	fr	tt	hin
1	0.636	0.680	0.679	0.717	0.512	0.679	0.612	0.746	0.713	0.388	0.565	0.695	0.577	0.484	0.721	0.455	0.456
2	0.603	0.711	0.673	0.676	0.501	0.540	0.601	0.677	0.694	0.355	0.542	0.680	0.572	0.486	0.752	0.371	0.389
3	0.589	0.712	0.673	0.679	0.570	0.530	0.549	0.669	0.664	0.255	0.511	0.641	0.553	0.485	0.741	0.363	0.285
4	0.656	0.724	0.712	0.748	0.539	0.682	0.631	0.773	0.730	0.369	0.607	0.728	0.647	0.499	0.801	0.485	0.480

Note: Strategy 1, 2, 3, and 4 correspond to single large model + prompt + word replacement, single large model + prompt + back-translation, single large model + prompt + translation detoxification, and multiple large models + prompt + word replacement, respectively. AvgP and AvgNP represent the average performance on primary and non-primary languages, respectively.

As table 6 shows our model outperforms most of the baseline methods in terms of Avgp score, including baseline_gpt4, baseline_o3mini, baseline_gpt4o, baseline_delete, baseline_backtranslation, and baseline_duplicate. Among the languages evaluated, Ukrainian (uk; 5th), Spanish (es; 3rd), and Hindi (hi; 2nd) achieved top-5 rankings in terms of performance. Furthermore, our AvNP score outperforms all baseline models and achieves 4th place overall in the test-phase evaluation. For this ranking , the top-performing languages are Japanese (ja; 4th), French (fr; 3rd), Hindi (hin; 4th), and Hebrew (he; 5th).

However, this method cannot completely solve the problem of homophones in different languages and cultures. For example, the English word "house" overlaps with the Chinese word "haosi", which means "good end". When this homophone appears in some toxic sentences, such as "You'll die a miserable death", this method cannot find the corresponding Chinese meaning. The sentence may be understood as "You won't have a good house".

Table 6
Final Grade In The Test

User	Avgp	en	es	de	zh	ar	hi	uk	ru	am	AvNP	it	ja	he	fr	tt	hin
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)	0.572 (12)	0.746 (8)	0.582 (13)	0.415 (23)	0.760 (9)	0.580 (4)	0.351 (18)
Jiaozipi	0.656 (8)	0.724 (8)	0.712 (3)	0.748 (9)	0.539 (9)	0.682 (9)	0.631 (2)	0.773 (5)	0.730 (9)	0.369 (21)	0.607 (4)	0.728 (11)	0.647 (4)	0.499 (5)	0.801 (3)	0.485 (16)	0.480 (4)
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)	0.579 (9)	0.742 (10)	0.637 (7)	0.513 (3)	0.780 (6)	0.468 (17)	0.333 (19)
baseline_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)
baseline_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)
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)	0.510 (21)	0.668 (21)	0.441 (26)	0.436 (18)	0.518 (29)	0.573 (7)	0.425 (9)
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)	0.342 (30)	0.462 (29)	0.241 (31)	0.339 (28)	0.626 (26)	0.254 (30)	0.133 (30)
baseline_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)

6. Summary

This paper briefly describes our work on the multilingual text detoxification task at PAN 2025. We propose using an ensemble of LLMs combined with prompt from the RISE framework to detoxify text across multiple languages. Initially, we constructed a toxicity-neutral text alignment dataset and a toxicity keyword list using the official dataset. Model inputs were created by integrating the RISE framework with few-shot methods. These inputs were used to drive multiple commercial LLMs (DeepSeek, Qwen, Kimi) to generate detoxified candidate outputs. Finally, the optimal output was selected through multi-dimensional evaluation, considering toxicity score, semantic integrity, and language fluency.For specific code, please refer to our release on github[8].

In this work, as shown in Table 6, the results demonstrate that our proposed method effectively handles the task of multilingual text detoxification, showing good adaptability and stability across different languages. However, the method does not adequately address homophones present in various languages and cultures. Future work will require more data for contextualization and research into frameworks for understanding homophones in LLMs. Additionally, we plan to enhance the tone restoration of detoxified text and construct a corresponding knowledge base to guide the result generation of LLMs.

[8]https://github.com/lxh44126/Detoxification/tree/code

7. Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62276064).

8. Declaration on Generative AI

During the preparation of this work, the authors used DouBao[9], YuanBao in order to: Grammar and spelling check, Paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Lu, B. Xu, X. Zhang, H. Wang, H. Zhu, D. Zhang, L. Yang, H. Lin, Towards comprehensive detection of chinese harmful memes, volume 37, Curran Associates, Inc., 2024, pp. 13302–13320.

[2] 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.

[3] 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.

[4] M. Guo, Z. Han, H. Chen, J. Peng, A Machine-Generated Text Detection Model Based on Text Multi-Feature Fusion, CEUR-WS.org, 2024, pp. 2593–2602.

[5] E. Řehulka, M. Šuppa, RAG Meets Detox: Enhancing Text Detoxification Using Open-Source Large Language Models with Retrieval Augmented Generation, CEUR-WS.org, 2024, pp. 3021–3031.

[6] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.

[7] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, volume 35, 2022, pp. 24824–24837.

[8] H. Zhang, X. Zhang, H. Huang, L. Yu, Prompt-based meta-learning for few-shot text classification, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 1342–1357.

[9] R. Dey, R. Mathur, Ensemble learning method using stacking with base learner, a comparison, Springer, Singapore, 2023, pp. 181–192.

[10] D. Dementieva, N. Babakov, A. Ronen, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. A. Moskovskiy, E. Stakovskii, E. Kaufman, A. Elnagar, A. Mukherjee, A. Panchenko, Multilingual and explainable text detoxification with parallel corpora, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 7998–8025.

[11] 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, CEUR-WS.org, 2024.

[12] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korencic, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova,

[9] https://www.doubao.com/chat/

E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative AI authorship verification - extended abstract, volume 14613 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 3–10.

[13] 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, CEUR-WS.org, 2024.

[14] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korencic, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative AI authorship verification - extended abstract, volume 14613 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 3–10.

[15] D. Dementieva, N. Babakov, A. Ronen, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. A. Moskovskiy, E. Stakovskii, E. Kaufman, A. Elnagar, A. Mukherjee, A. Panchenko, Multilingual and explainable text detoxification with parallel corpora, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 7998–8025.