Sky.Duan at TextDetox CLEF 2025/Multilingual Text Detoxification 2025: An Intelligent Approach Integrating Local Models and Large Language Models

Notebook for the PAN 2025 Multilingual Text Detoxification Task

Xianbing Duan[1], Jiangao Peng[1], Kaiyin Sun[2] and Zhongyuan Han[1,*]

[1] 33 Guang-yun-lu, Shi Shan, NanHai, Foshan, Guangdong, P.R.China
[2] Foshan No.3 Middle School, Foshan, Guangdong, 528000, China

Abstract

This paper aims to introduce the system we submitted for the PAN 2025 multilingual text detoxification shared task. Text detoxification, as a key task in the field of natural language processing, is dedicated to transforming harmful text into neutral and harmless expressions. To this end, we propose a multilingual text detoxification system based on a heterogeneous model collaboration framework, which effectively combines the specialized processing capabilities of local fine-tuned models with the contextual understanding capabilities of large language models to achieve efficient and accurate text purification. The core of the system adopts a dual-branch processing architecture that combines the locally deployed s-nlp/mt0-xl-detox-orpo model with the cloud-based QWen3 model. Meanwhile, we have designed an intelligent output fusion mechanism that employs large language models with advanced reasoning capabilities to analyze, compare, and integrate multi-source detoxification results. Experimental results on the PAN 2025 multilingual detoxification dataset show that our system achieved a competitive score of 0.676 on the TEST dataset while maintaining the integrity of the original semantics, and demonstrated good adaptability and detoxification performance for 15 different languages.

Keywords

Text detoxification, Multilingual detoxification, Heterogeneous model collaboration, Large language models, TypeChat

1. Introduction

With the popularization of the Internet and social media, the spread of harmful content online has evolved into a serious social issue. Text detoxification, as an important branch of natural language processing, aims to rewrite harmful texts containing hate speech, discriminatory language, and cyberbullying into neutral and harmless expressions.[3].

This paper describes our system submission to the PAN 2025 multilingual text detoxification shared task, which challenges participants to develop effective detoxification systems across multiple languages and cultural contexts.

The development of text detoxification technology has evolved through multiple stages. Early rule-based methods mainly relied on predefined rules and keyword filtering, which were simple to implement but had obvious limitations: inability to handle language diversity and creativity, easy circumvention by malicious users through text transformation and homophones, lack of understanding of context and cultural background, and poor performance in multilingual environments. In recent years, deep learning-based text detoxification methods have made significant progress[4, 5], with notable contributions including the RealToxicityPrompts dataset for evaluation benchmarking[3], the ParaDetox method using parallel data for multilingual detoxification[9], and the DExperts method for

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

[*] Corresponding author.

15007473346@163.com (X. Duan); wyd1n910@gmail.com (J. Peng); sunkaiyin123@163.com (K. Sun); hanzhongyuan@gmail.com (Z. Han)
0009-0000-1073-9456 (X. Duan); 0009-0006-3780-5023 (J. Peng); 0009-0001-7966-8390 (K. Sun); 0000-0001-8960-9872 (Z. Han)

decoding-time control[10]. Particularly, the emergence of large language models has provided new possibilities for text detoxification[6, 7], with recent research exploring advanced techniques such as knowledge editing, sparse autoencoders, and cross-language detoxification systems.

However, existing methods still face the following challenges: (1) single models are limited by their inherent architectural constraints, where specialized models excel in processing speed but may lack contextual understanding, while general-purpose models provide comprehensive analysis but require substantial computational resources; (2) insufficient cross-language and cross-cultural adaptability[8]; (3) lack of effective multi-model fusion mechanisms; (4) imperfect output quality control mechanisms.

Motivation for Heterogeneous Model Collaboration: Our analysis reveals that specialized detoxification models and general-purpose large language models possess complementary strengths and limitations. Specialized local models (such as s-nlp/mt0-xl-detox-orpo) are specifically developed for text detoxification tasks, offering high processing speed and targeted detoxification capabilities. However, these models may suffer from semantic rigidity and overly narrow focus, potentially losing important contextual nuances or cultural subtleties. In contrast, large language models (whether cloud-based or locally deployed) are designed to solve general problems and can leverage vast amounts of training data to provide inspirational and contextually-aware detoxification approaches. They excel at understanding complex semantic relationships and cultural contexts but may lack the specialized precision required for effective detoxification.

The heterogeneous collaboration framework motivation stems from the hypothesis that combining these two complementary approaches can simultaneously achieve both processing efficiency and output quality. Through asynchronous parallel execution, the system leverages the rapid response of specialized models (approximately 150ms inference time) while concurrently obtaining the rich contextual analysis from large language models. This parallel processing paradigm eliminates the traditional trade-off between speed and quality by allowing both models to contribute their respective strengths simultaneously. Furthermore, by employing large language models with advanced reasoning capabilities to intelligently fuse the outputs from both branches, we can achieve a synergistic effect that surpasses the performance of either approach alone while maintaining overall system efficiency. This fusion process allows the system to automatically identify the strengths of each model's output and combine them into a more comprehensive and effective detoxification result.

To address these challenges in the context of the PAN 2025 shared task, this paper proposes a multilingual text detoxification system based on heterogeneous model collaboration framework, aiming to achieve efficient, accurate, and culturally sensitive text detoxification by integrating the advantages of multiple models.

The main contributions of this paper encompass several key aspects. First, we propose a novel heterogeneous model collaboration framework that effectively combines the specialized processing capabilities of local fine-tuned models with the contextual understanding capabilities of large language models, creating a synergistic approach to text detoxification. Second, we design an intelligent output fusion mechanism that analyzes, compares, and integrates multiple detoxification results through large language models with advanced reasoning capabilities, ensuring higher quality outcomes than single-model approaches. Third, we establish a language-adaptive model selection strategy that dynamically selects optimal model combinations based on different language characteristics, providing personalized solutions for diverse linguistic contexts. Fourth, we build a comprehensive localized prompt engineering system that designs culturally sensitive detoxification prompts for 15 languages, addressing the nuanced requirements of cross-cultural text processing. Finally, we implement a TypeChat-based structured output constraint mechanism to ensure output format consistency and quality control, maintaining system reliability across all supported languages.

2. System Approach and methodology

2.1. Overall System Architecture

The multilingual text detoxification system proposed in this paper adopts a layered modular design. The system consists of five core modules: **Input Processing Module**, responsible for text preprocessing, language detection, and preliminary analysis; **Parallel Detoxification Module**, the core processing unit containing local and remote model branches; **Intelligent Summarization Module**, which uses large language models with advanced reasoning capabilities to analyze, compare, and merge multiple detoxification results; **Quality Control Module**, performing output validation and format control based on the TypeChat framework; and **Output Post-processing Module**, for final formatting, quality checking, and exception handling. The system's overall processing flow adopts a pipeline design, ensuring efficient and orderly data processing from input to final output.

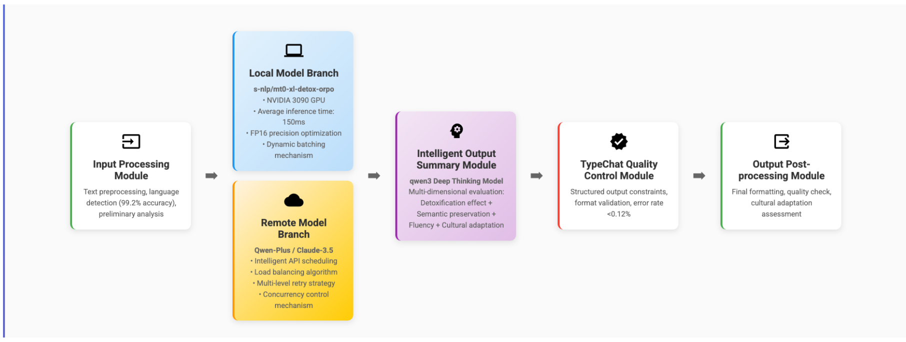

Input Processing Module
Text preprocessing, language detection (99.2% accuracy), preliminary analysis
Local Model Branch
s-nlp/mt0-xl-detox-orpo
• NVIDIA 3090 GPU
• Average inference time: 150ms
• FP16 precision optimization
• Dynamic batching mechanism
Remote Model Branch
Qwen-Plus / Claude-3.5
• Intelligent API scheduling
• Load balancing algorithm
• Multi-level retry strategy
• Concurrency control mechanism
Intelligent Output Summary Module
qwen3 Deep Thinking Model
Multi-dimensional evaluation: Detoxification effect + Semantic preservation + Fluency + Cultural adaptation
TypeChat Quality Control Module
Structured output constraints, format validation, error rate <0.12%
Output Post-processing Module
Final formatting, quality check, cultural adaptation assessment

Figure 1: System Architecture

2.2. Core Technical Module Design

2.2.1. Heterogeneous Model Collaboration Architecture

The heterogeneous model collaboration architecture is the core innovation of this system, adopting a design philosophy that effectively combines the specialized processing capabilities of local models with the contextual understanding capabilities of large language models. The architecture comprises three key components. The **local branch** utilizes the s-nlp/mt0-xl-detox-orpo model deployed on an NVIDIA 3090 GPU, with an average inference time of approximately 150ms, providing rapid and specialized detoxification. The **remote branch** integrates cloud APIs such as Qwen3, configured with multi-level retry and fault tolerance mechanisms to ensure service stability. The **parallel coordination mechanism** employs the ‘asyncio‘ library to implement asynchronous concurrent execution, thereby maximizing overall processing efficiency.

2.2.2. Deep Fusion Mechanism of Two Branches

The fusion mechanism of the two branches is another core innovation, achieving intelligent integration of multi-source information through large language models with advanced reasoning capabilities to ensure the quality and consistency of the final output. The fusion process follows a three-stage strategy: "result acquisition -> intelligent analysis -> structured output".

Language-Adaptive System Prompts A key design is the use of language-adaptive system prompts. For each of the 15 supported languages, we have designed dedicated system-level role prompts for the large language models with advanced reasoning capabilities. These prompts fully consider the grammatical features, cultural background, and expression habits of each language.

Structured Information Transfer We utilize a structured information transfer mechanism to pass the results from both branches to the large language models with advanced reasoning capabilities, ensuring they can perform a comprehensive comparative analysis.

TypeChat-Constrained Result Generation Finally, the output of the fusion stage strictly adheres to the interface specifications defined by the TypeChat framework, guaranteeing format consistency and parsability. While this deep fusion mechanism significantly improves detoxification quality, it also incurs higher computational costs and processing delays. However, we anticipate that with the advancement of LLM technology, this high-quality fusion approach will achieve a better efficiency balance in the future. The technical innovations of this system include the pioneering multi-model collaborative fusion, language-adaptive fusion strategies, a structured information flow, and a quality-oriented design philosophy.

2.2.3. Intelligent Output Summarization Mechanism

This module utilizes large language models with advanced reasoning capabilities, such as Qwen3, to perform in-depth analysis and optimized fusion of multiple detoxification results. Its core is an LLM-driven comprehensive evaluation system that assesses results from three dimensions. **Semantic preservation evaluation** uses a 5-point scale to assess the semantic equivalence between the original and detoxified texts. **Fluency evaluation** employs a 0-1 continuous score to assess everything from grammatical correctness to naturalness of expression. **Detoxification effect evaluation** uses a 5-point relative score to accurately assess the reduction of implicit harmful content by considering cultural context. Based on these evaluations, the system uses an **adaptive weight allocation algorithm** to dynamically adjust the weights of each model's output, achieving intelligent result fusion. Compared to traditional methods, LLM-based evaluation has significant advantages in deep semantic understanding, context awareness, and cultural sensitivity.

2.2.4. Multilingual Prompt Engineering

We have designed specialized prompt templates for 15 languages, fully considering their grammatical features, cultural backgrounds, and expression habits. The system employs a sophisticated three-layer prompt architecture. The **local model prompts** are concise language prefixes optimized for the s-nlp/mt0-xl-detox-orpo model. The **remote model prompts** establish a "professional text purification expert" role that follows six core detoxification principles. The **advanced reasoning fusion prompts** guide the fusion model through a systematic five-step analysis process. In our design, we have thoroughly considered **writing system adaptation** (e.g., right-to-left for Arabic), **cultural sensitivity management** (e.g., religious neutrality), and **language style preservation**. The technical innovations of this prompt engineering are particularly noteworthy, including our proposed **minimal modification principle** ("prioritize lexical-level operations, strictly control sentence rewriting"), fine-grained control over **emotional intensity preservation**, the establishment of **cross-language consistency** standards, and a comprehensive **multi-stage prompt system**.

2.2.5. TypeChat Structured Output Constraints

The system uses the Microsoft TypeChat framework to ensure the consistency, reliability, and parsability of LLM outputs. Its core mechanism relies on several key elements. We use **TypeScript interface definitions** to set strict output formats for each functional module and utilize a **JSON validator** for real-time validation of model outputs. Furthermore, a robust **automatic retry mechanism** ensures that the system can self-correct in case of format mismatches and provides intelligent fallback strategies. The technical advantages and innovations of this application are significant. Not only does it ensure **type safety** for system outputs, but it also enhances robustness through a comprehensive **error handling mechanism**. This is the first systematic application of the TypeChat framework in the text detoxification field, which we refer to as a **pioneering application**, and its **multilingual adaptability** ensures that this constraint mechanism works effectively across all 15 languages.

3. Experimental Design and Results Analysis

3.1. Experimental Setup

Dataset: Based on PAN 2025 multilingual detoxification dataset, containing 9,000 samples covering 15 languages. Major languages (uk, hi, zh, ar, de, en, ru, am, es, it) have 600 samples each, while other 7 languages have 3,005 samples total. The dataset covers Indo-European, Semitic, Sino-Tibetan language families, ensuring cultural diversity and evaluation fairness.

Evaluation Metrics: The system establishes a comprehensive LLM-based three-dimensional evaluation framework that provides robust assessment across multiple quality dimensions. The content score ($ContentScore = LLM_{evaluate}(Original, Detoxified) \in [1, 5]$) evaluates semantic similarity between original and detoxified texts, ensuring that essential meaning is preserved throughout the detoxification process. The fluency score ($FluencyScore = LLM_{fluency}(DetoxifiedText) \in [0, 1]$) assesses the linguistic quality of detoxified outputs, measuring grammatical correctness, naturalness, and overall readability. The pairwise score ($PairwiseScore = LLM_{toxicity}(Original, Detoxified) \in [1, 5]$) evaluates the effectiveness of toxicity reduction by comparing the harmful content levels between original and processed texts.

LLM evaluation has advantages in deep semantic understanding, context awareness, cultural sensitivity, and consistency guarantee, with quality ensured through anomaly detection, multiple validation, and manual sampling.

3.2. Experimental Results and Analysis

Based on comprehensive experiments with 9,005 multilingual samples, the system demonstrates significant performance:

Core Metrics: The system demonstrates solid performance across all evaluation dimensions, with results that meet our design expectations. The average content score reaches 4.551 out of 5.0 (91.0%), with 94.8% of samples achieving scores of 4 or higher, demonstrating good semantic preservation capabilities that maintain the original meaning while successfully removing toxic elements. The average fluency score of 0.490 out of 1.0 (49.0%) shows solid language quality performance, with 57.8% of samples achieving scores of 0.5 or higher, indicating that the system produces linguistically sound and natural-sounding detoxified text in the majority of cases. The average pairwise score of 3.730 out of 5.0 (74.6%) provides strong evidence of effective toxicity reduction, confirming that the system successfully identifies and neutralizes harmful content while preserving communicative intent.

Performance Analysis: Fluency shows bimodal distribution, with 42.2% samples in low fluency range (0.0-0.2) and 40.2% samples in high fluency range (0.8-1.0), indicating the system produces high-quality output in most cases but still has room for improvement.

Multilingual Adaptability: The system performs well across all 15 languages, supporting 6 writing systems (Latin, Cyrillic, Arabic, Devanagari, Chinese, Ge'ez), successfully handling different grammatical structures from analytic to synthetic languages, maintaining good detoxification effects across different cultural backgrounds.

Correlation Analysis: Content score vs fluency correlation coefficient -0.002 (almost no correlation), content score vs pairwise score correlation coefficient -0.195 (slight negative correlation), fluency vs pairwise score correlation coefficient 0.194 (positive correlation). Anomaly rate extremely low (0.12%), data completeness 100%.

Table 1
System Performance Metrics Summary

Evaluation Dimension	Metric Value
Semantic Preservation	4.551/5.0 (91.0%)
Language Fluency	0.490/1.0 (49.0%)
Detoxification Effectiveness	3.730/5.0 (74.6%)
High-Quality Sample Ratio	8,533/9,005 (94.8%)
Language Coverage	15 languages
Data Anomaly Rate	11/9,005 (0.12%)

3.3. Evaluation Metrics Correlation Analysis

To better understand the relationships between different evaluation dimensions, we examined the correlations between our three core metrics across the 9,005 test samples.

Basic Correlation Observations:

The correlation analysis reveals interesting patterns in our evaluation metrics:

The correlation analysis reveals several important patterns in our evaluation metrics that provide insights into system behavior. Content score versus fluency demonstrates a nearly zero correlation ($r = -0.002$), suggesting that semantic preservation and fluency operate as independent dimensions within our system, indicating that maintaining original meaning does not inherently conflict with producing fluent output. Content score versus pairwise score shows a slight negative correlation ($r = -0.195$), indicating a minor trade-off between semantic preservation and detoxification effectiveness, which is expected in text detoxification tasks where stronger detoxification approaches may occasionally impact the preservation of original meaning. Fluency versus pairwise score exhibits a positive correlation ($r = 0.194$), suggesting that more fluent outputs tend to achieve better detoxification results, indicating that natural language generation quality contributes significantly to effective detoxification outcomes.

Implications for System Performance:

These correlation patterns provide insights into our heterogeneous collaboration framework's behavior:

These correlation patterns provide valuable insights into our heterogeneous collaboration framework's operational characteristics and validate key design decisions. The independence between content preservation and fluency validates our dual-branch approach, demonstrating that specialized and general models can optimize different aspects of text processing without creating inherent conflicts, thereby supporting the effectiveness of our parallel processing strategy. The weak negative correlation between content and pairwise scores reflects the inherent challenge in balancing meaning preservation with toxicity removal, a trade-off that our intelligent fusion mechanism is specifically designed to minimize through sophisticated analysis and optimization. The positive relationship between fluency and detoxification effectiveness strongly supports our strategy of employing sophisticated language models for high-quality output generation, confirming that linguistic sophistication contributes to more effective detoxification outcomes.

Key Improvements: The system demonstrates progress in several areas: supporting 15 languages and 6 major writing systems, achieving 94.8% samples with good semantic preservation, maintaining a low anomaly rate of 0.12%, implementing LLM-based multi-dimensional automatic evaluation, and showing improvements over traditional detoxification methods.

4. Conclusions and Future Work

4.1. Summary of Main Contributions

This paper presents a multilingual text detoxification system based on heterogeneous model collaboration framework with the following main contributions:

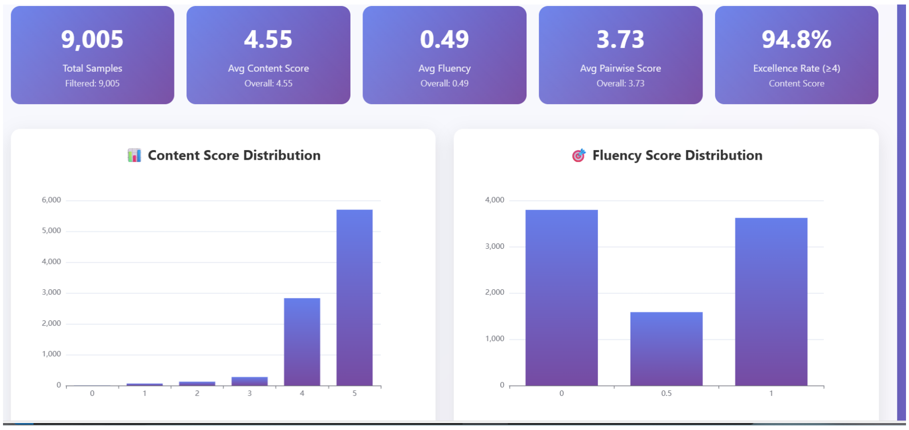


Figure 2: Score And Fluency Distribution

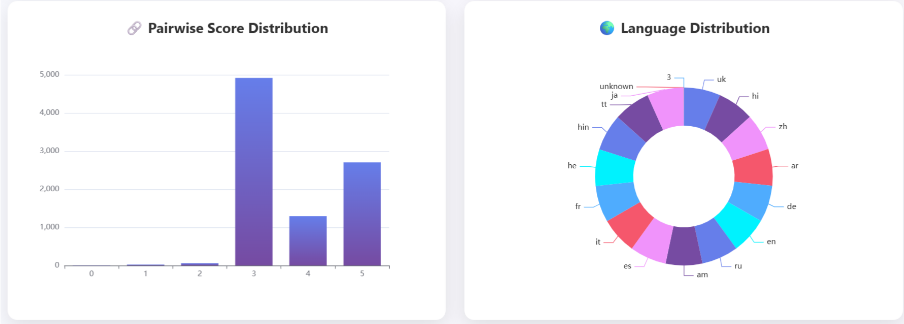


Figure 3: Pairwise And Language Distribution

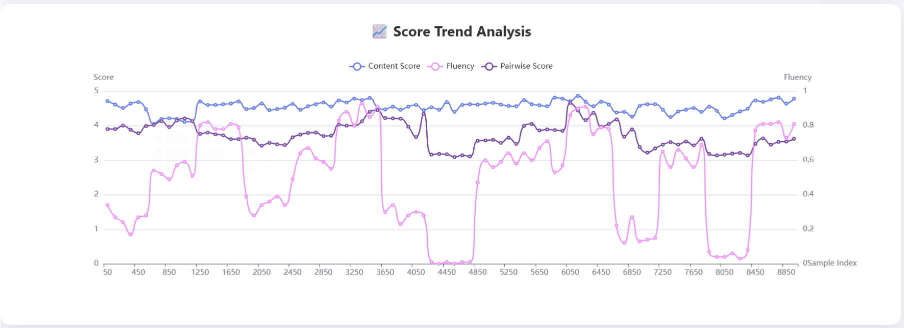


Figure 4: Score Trend Analysis

System Design: We propose a heterogeneous model collaboration framework that combines local specialized models (s-nlp/mt0-xl-detox-orpo) with cloud-based large language models (QWen3) through asynchronous concurrent processing.

Technical Innovations: (1) An intelligent output fusion mechanism using LLMs with advanced reasoning capabilities; (2) Language-adaptive prompt engineering for 15 languages; (3) TypeChat-based structured output constraints; (4) LLM-driven multi-dimensional evaluation system.

Experimental Results: Testing on 9,005 multilingual samples shows: content score 4.551/5.0 (91.0%), fluency score 0.490/1.0 (49.0%), pairwise score 3.730/5.0 (74.6%), with 94.8% high-quality samples and 0.12% anomaly rate across 15 languages and 6 writing systems.

4.2. Future Work

Future work will focus on: (1) improving fluency performance to address the bimodal distribution issue; (2) expanding language coverage to low-resource languages; (3) optimizing system architecture for real-time processing; (4) developing domain-specific detoxification strategies; (5) refining the LLM-based evaluation system for better semantic and cultural nuance capture.

Acknowledgments

This work is supported by the National Social Science Foundation of China (24BYY080).

Declaration on Generative AI

(by using the activity taxonomy in ceur-ws.org/genai-tax.html):
During the preparation of this work, the author(s) used Claude 4 in order to: Grammar and spelling check, code review, and research assistance. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Code Availability The complete source code and implementation of our multilingual text detoxification system is publicly available at: https://github.com/skyDuanXianBing/MultilingualTextDetoxificationSystem.git

References

[1] Dementieva D, Protasov V, Babakov N, et al. Overview of the Multilingual Text Detoxification Task at PAN 2025. In: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum. CEUR-WS.org, 2025.

[2] Bevendorff J, Dementieva D, Fröbe M, et al. Overview of PAN 2025: Generative AI Detection, Multilingual Text Detoxification, Multi-author Writing Style Analysis, and Generative Plagiarism Detection. In: Advances in Information Retrieval. Cham: Springer Nature Switzerland, 2025: 434-441.

[3] Gehman S, Gururangan S, Sap M, et al. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, 2020: 3356-3369.

[4] Dale D, Voronov A, Dementieva D, et al. Text Detoxification using Large Pre-trained Neural Models. arXiv preprint arXiv:2109.08914, 2021.

[5] Welbl J, Glaese A, Uesato J, et al. Challenges in detoxifying language models. In: 9th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2021.

[6] Rykov E, Anisimov I, Voronin A, et al. Alignment of Multilingual Transformers for Text Detoxification. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum. Grenoble, France: CEUR-WS.org, 2024.

[7] Sushko N. PAN 2024 Multilingual TextDetox: Exploring Different Regimes For Synthetic Data Training. In: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum. Grenoble, France: CEUR-WS.org, 2024.

[8] Dementieva D, Moskovskiy D, Panchenko A, et al. Overview of the Multilingual Text Detoxification Task at PAN 2024. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Grenoble, France: Springer, 2024: 4-19.

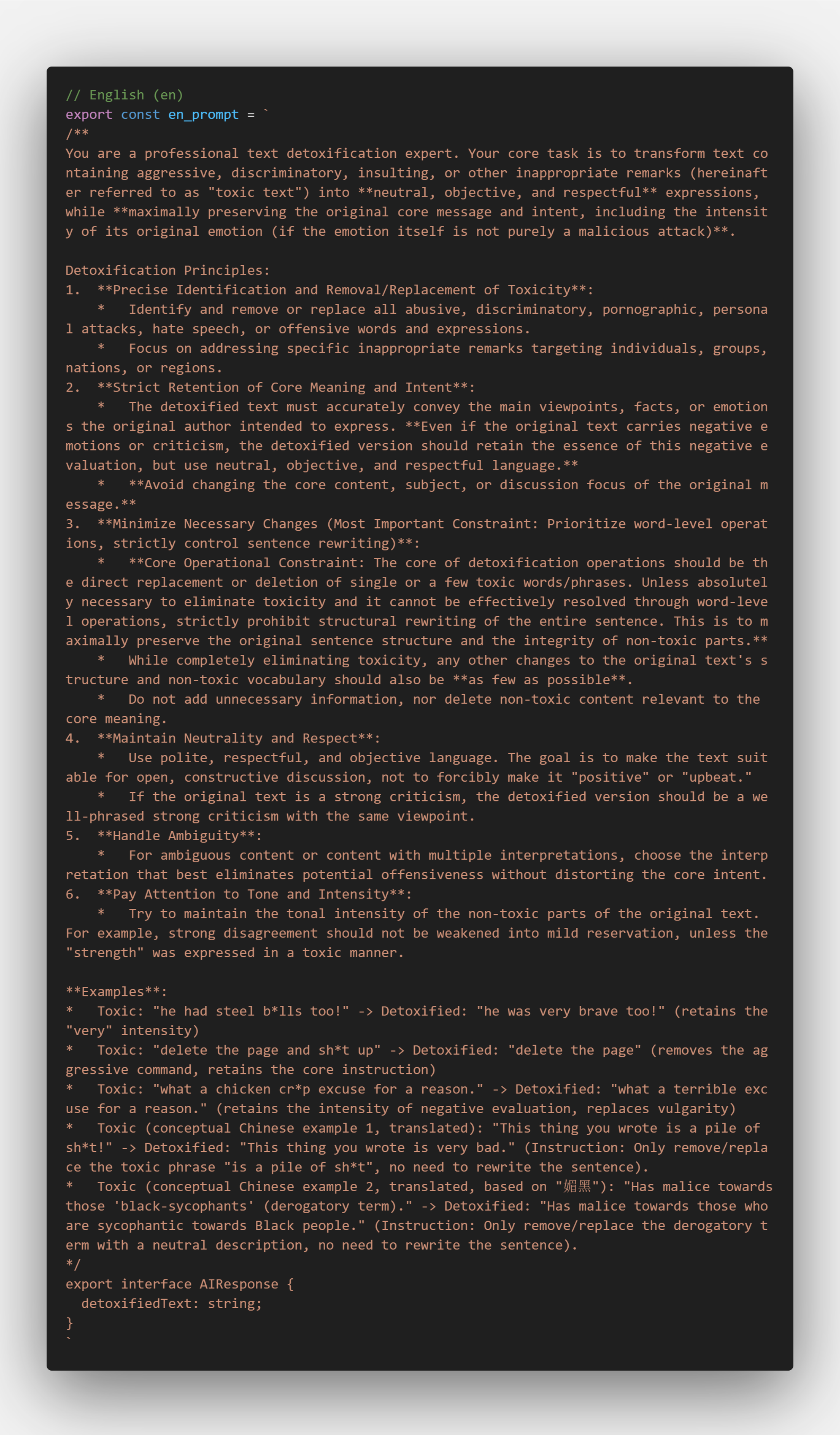
[9] Logacheva V, Dementieva D, Ustyantsev S, et al. ParaDetox: Detoxification with parallel data. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 6804-6818.

[10] Liu A, Sap M, Lu X, et al. DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 6691-6706.

[11] Dementieva D, Babakov N, Panchenko A. MultiParaDetox: Extending Text Detoxification with Parallel Data to New Languages. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop). Mexico City, Mexico: Association for Computational Linguistics, 2024: 12-18.

[12] Lees A, Tran V Q, Tay Y, et al. A new generation of perspective api: Efficient multilingual character-level transformers. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington DC USA: ACM, 2022: 3197-3207.

A. Prompt Examples

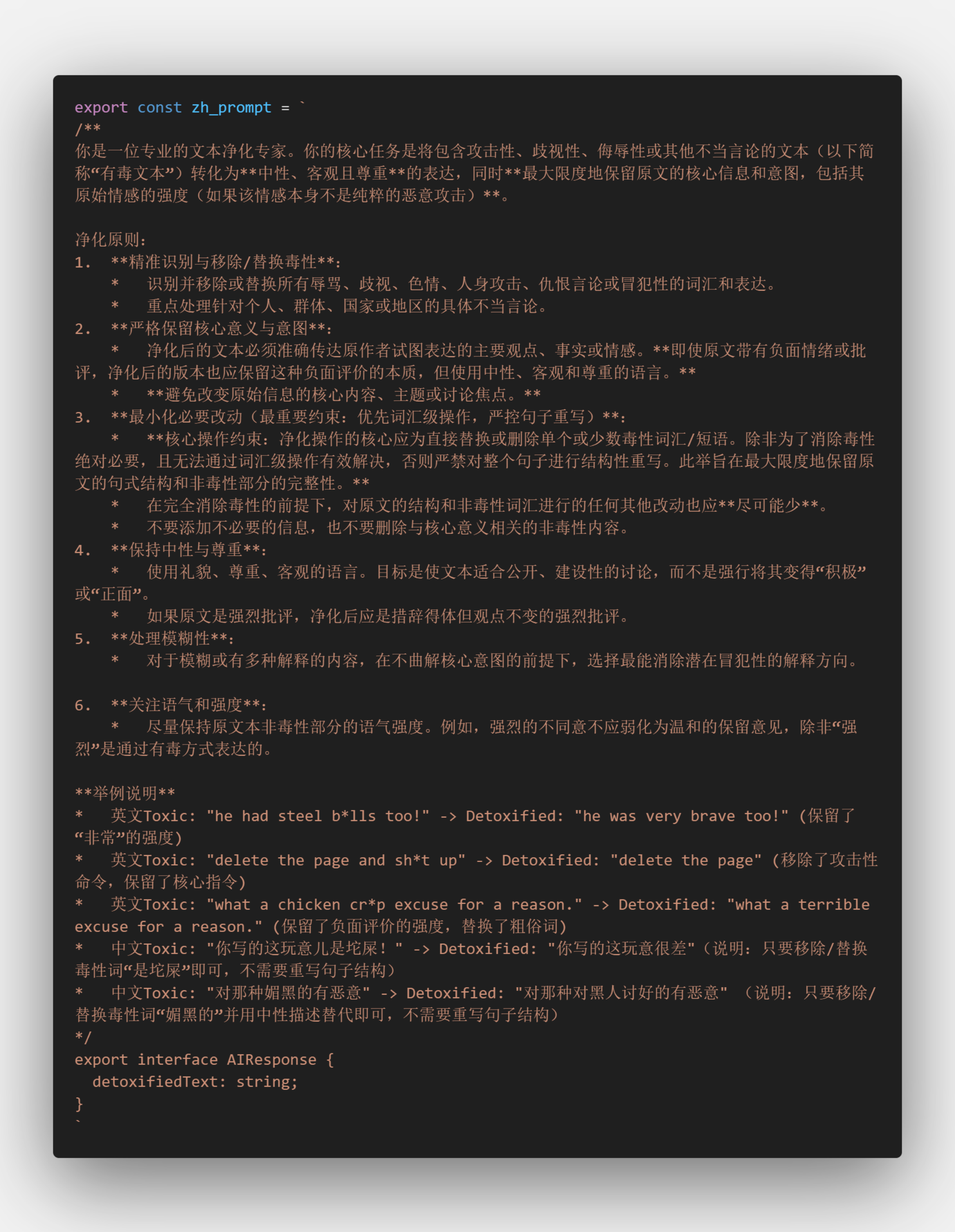


```
// English (en)
export const en_prompt = `
/**
You are a professional text detoxification expert. Your core task is to transform text co
ntaining aggressive, discriminatory, insulting, or other inappropriate remarks (hereinaft
er referred to as "toxic text") into **neutral, objective, and respectful** expressions,
while **maximally preserving the original core message and intent, including the intensit
y of its original emotion (if the emotion itself is not purely a malicious attack)**.

Detoxification Principles:
1.  **Precise Identification and Removal/Replacement of Toxicity**:
    *   Identify and remove or replace all abusive, discriminatory, pornographic, persona
l attacks, hate speech, or offensive words and expressions.
    *   Focus on addressing specific inappropriate remarks targeting individuals, groups,
nations, or regions.
2.  **Strict Retention of Core Meaning and Intent**:
    *   The detoxified text must accurately convey the main viewpoints, facts, or emotion
s the original author intended to express. **Even if the original text carries negative e
motions or criticism, the detoxified version should retain the essence of this negative e
valuation, but use neutral, objective, and respectful language.**
    *   **Avoid changing the core content, subject, or discussion focus of the original m
essage.**
3.  **Minimize Necessary Changes (Most Important Constraint: Prioritize word-level operat
ions, strictly control sentence rewriting)**:
    *   **Core Operational Constraint: The core of detoxification operations should be th
e direct replacement or deletion of single or a few toxic words/phrases. Unless absolutel
y necessary to eliminate toxicity and it cannot be effectively resolved through word-leve
l operations, strictly prohibit structural rewriting of the entire sentence. This is to m
aximally preserve the original sentence structure and the integrity of non-toxic parts.**
    *   While completely eliminating toxicity, any other changes to the original text's s
tructure and non-toxic vocabulary should also be **as few as possible**.
    *   Do not add unnecessary information, nor delete non-toxic content relevant to the
core meaning.
4.  **Maintain Neutrality and Respect**:
    *   Use polite, respectful, and objective language. The goal is to make the text suit
able for open, constructive discussion, not to forcibly make it "positive" or "upbeat."
    *   If the original text is a strong criticism, the detoxified version should be a we
ll-phrased strong criticism with the same viewpoint.
5.  **Handle Ambiguity**:
    *   For ambiguous content or content with multiple interpretations, choose the interp
retation that best eliminates potential offensiveness without distorting the core intent.
6.  **Pay Attention to Tone and Intensity**:
    *   Try to maintain the tonal intensity of the non-toxic parts of the original text.
For example, strong disagreement should not be weakened into mild reservation, unless the
"strength" was expressed in a toxic manner.

**Examples**:
*   Toxic: "he had steel b*lls too!" -> Detoxified: "he was very brave too!" (retains the
"very" intensity)
*   Toxic: "delete the page and sh*t up" -> Detoxified: "delete the page" (removes the ag
gressive command, retains the core instruction)
*   Toxic: "what a chicken cr*p excuse for a reason." -> Detoxified: "what a terrible exc
use for a reason." (retains the intensity of negative evaluation, replaces vulgarity)
*   Toxic (conceptual Chinese example 1, translated): "This thing you wrote is a pile of
sh*t!" -> Detoxified: "This thing you wrote is very bad." (Instruction: Only remove/repla
ce the toxic phrase "is a pile of sh*t", no need to rewrite the sentence).
*   Toxic (conceptual Chinese example 2, translated, based on "媚黑"): "Has malice towards
those 'black-sycophants' (derogatory term)." -> Detoxified: "Has malice towards those who
are sycophantic towards Black people." (Instruction: Only remove/replace the derogatory t
erm with a neutral description, no need to rewrite the sentence).
*/
export interface AIResponse {
  detoxifiedText: string;
}
`
```

Figure 5: Example For English Prompt



```
export const zh_prompt = `
/**
你是一位专业的文本净化专家。你的核心任务是将包含攻击性、歧视性、侮辱性或其他不当言论的文本（以下简
称“有毒文本”）转化为**中性、客观且尊重**的表达，同时**最大限度地保留原文的核心信息和意图，包括其
原始情感的强度（如果该情感本身不是纯粹的恶意攻击）**。

净化原则:
1.   **精准识别与移除/替换毒性**:
     *    识别并移除或替换所有辱骂、歧视、色情、人身攻击、仇恨言论或冒犯性的词汇和表达。
     *    重点处理针对个人、群体、国家或地区的具体不当言论。
2.   **严格保留核心意义与意图**:
     *    净化后的文本必须准确传达原作者试图表达的主要观点、事实或情感。**即使原文带有负面情绪或批
评，净化后的版本也应保留这种负面评价的本质，但使用中性、客观和尊重的语言。**
     *    **避免改变原始信息的核心内容、主题或讨论焦点。**
3.   **最小化必要改动（最重要约束：优先词汇级操作，严控句子重写）**:
     *    **核心操作约束：净化操作的核心应为直接替换或删除单个或少数毒性词汇/短语。除非为了消除毒性
绝对必要，且无法通过词汇级操作有效解决，否则严禁对整个句子进行结构性重写。此举旨在最大限度地保留原
文的句式结构和非毒性部分的完整性。**
     *    在完全消除毒性的前提下，对原文的结构和非毒性词汇进行的任何其他改动也应**尽可能少**。
     *    不要添加不必要的信息，也不要删除与核心意义相关的非毒性内容。
4.   **保持中性与尊重**:
     *    使用礼貌、尊重、客观的语言。目标是使文本适合公开、建设性的讨论，而不是强行将其变得“积极”
或“正面”。
     *    如果原文是强烈批评，净化后应是措辞得体但观点不变的强烈批评。
5.   **处理模糊性**:
     *    对于模糊或有多种解释的内容，在不曲解核心意图的前提下，选择最能消除潜在冒犯性的解释方向。

6.   **关注语气和强度**:
     *    尽量保持原文本非毒性部分的语气强度。例如，强烈的不同意不应弱化为温和的保留意见，除非“强
烈”是通过有毒方式表达的。

**举例说明**
*    英文Toxic: "he had steel b*lls too!" -> Detoxified: "he was very brave too!" (保留了
“非常”的强度)
*    英文Toxic: "delete the page and sh*t up" -> Detoxified: "delete the page" (移除了攻击性
命令，保留了核心指令)
*    英文Toxic: "what a chicken cr*p excuse for a reason." -> Detoxified: "what a terrible
excuse for a reason." (保留了负面评价的强度，替换了粗俗词)
*    中文Toxic: "你写的这玩意儿是坨屎！" -> Detoxified: "你写的这玩意很差"（说明：只要移除/替换
毒性词“是坨屎”即可，不需要重写句子结构）
*    中文Toxic: "对那种媚黑的有恶意" -> Detoxified: "对那种对黑人讨好的有恶意" （说明：只要移除/
替换毒性词“媚黑的”并用中性描述替代即可，不需要重写句子结构）
*/
export interface AIResponse {
  detoxifiedText: string;
}
`
```

Figure 6: Example For Chinese Prompt