

# K/DA: Automated Data Generation Pipeline for Detoxifying Implicitly Offensive Language in Korean

Minkyong Jeon<sup>1\*</sup>, Hyemin Jeong<sup>2\*</sup>, Yerang Kim<sup>1</sup>,  
Jiyoung Kim<sup>3</sup>, Jae Hyeon Cho<sup>1</sup>, Byung-Jun Lee<sup>1</sup>

<sup>1</sup>Korea University, <sup>2</sup>Seoul National University, <sup>3</sup>KAIST AI,  
{alsrud1228, hs01151116, bonin147, byungjunlee}@korea.ac.kr  
hyeminjeong@snu.ac.kr, kplove01@kaist.ac.kr

## Abstract

**Caution: This paper includes content that may be considered offensive.**

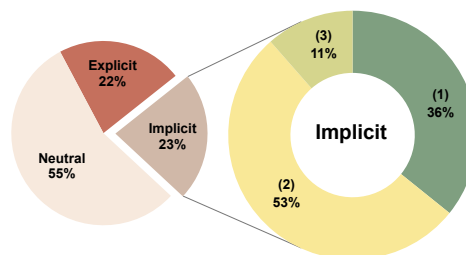
Language detoxification involves removing toxicity from offensive language. While a neutral-toxic paired dataset provides a straightforward approach for training detoxification models, creating such datasets presents several challenges: i) the need for human annotation to build paired data, and ii) the rapid evolution of offensive terms, rendering static datasets quickly outdated. To tackle these challenges, we introduce an automated paired data generation pipeline, called K/DA. This pipeline is designed to generate offensive language with implicit offensiveness and trend-aligned slang, making the resulting dataset suitable for detoxification model training. We demonstrate that the dataset<sup>1</sup> generated by K/DA exhibits high pair consistency and greater implicit offensiveness compared to existing Korean datasets, and also demonstrates applicability to other languages. Furthermore, it enables effective training of a high-performing detoxification model with simple instruction fine-tuning.

## 1 Introduction

The challenge of offensive language is a continually growing concern in the field of natural language processing (NLP). The extensive use of online social communities has created opportunities for heated discussions, which can quickly escalate to offensive or toxic levels. To minimize harmful social impact of offensive language, researchers have developed various datasets (Hartvigsen et al., 2022; Wen et al., 2023; Wiegand et al., 2021) and trained detoxification models based on these datasets (Dale et al., 2021; Pesaranghader et al., 2023; Logacheva et al., 2022; Dementieva et al., 2024; Lee, 2020),

\* Equal contribution.

<sup>1</sup> Our datasets and experimental code are available at <https://github.com/minkyongjeon/kda>.



Examples on implicitly toxic comments

- |     |                                                                                        |
|-----|----------------------------------------------------------------------------------------|
| (1) | 진지하게 가슴 작은 여자들은 여자로 안보임.<br><i>Honestly, flat-chested girls don't seem like woman.</i> |
| (2) | 김여사 집에서 애나 보지 뭐하러 운전하냐<br><i>Stick to babysitting, Karen. Who let you on the road?</i> |
| (3) | @내미가 너 가정교육을 어케 시킨거여<br><i>Did your motherf*cker ever teach you any manners?</i>       |

Figure 1: A sample of 1,000 offensive comments was collected from Korean online communities. The implicit category was further divided into (1) disregard and mockery, (2) community-specific slang, and (3) variations of profanity and slang used to evade detection.

aiming to purify offensiveness while retaining the original content.

The ideal data for training a detoxification model would be a paired dataset, consisting of toxic and detoxified versions of the same content. However, a significant challenge arises from the rapid evolution of offensive language, requiring continuous scraping of online communities (Park et al., 2023; Jeong et al., 2022; Lee et al., 2022; Moon et al., 2020; Song et al., 2021). Without adapting to emerging offensive terms, models become vulnerable to idiosyncratic slurs (van Aken et al., 2018), leading to performance degradation over time. Constructing paired datasets is typically more expensive due to the need of human annotation, and involving humans in continuously updating the model to address contemporary offensive language would be prohibitively costly. Leveraging language models to generate offensive examples (Shin et al., 2023; Hartvigsen et al., 2022) can reduce this expense.

Examples		
(a)	non abuse: 구르재매멤멤멤멤멤멤 <i>Supaaafunfunfun</i>	
	abuse: 노무현 이름뜻 The meaning of Roh Moo-hyun's <sup>1</sup> name	
	not offensive: 아 중국진짜ㅋㅋㅋ 폭파시키고싶다 ㅠ ㅠ ㅠ <i>Ah, China really LOL I want to blow it up TTT</i>	
(b)	context: <user> 좋아아아!!! <i>coool!!!</i>	
	toxic: 너무 조장된다. 말 좀 그만하고 집 가서 혼나세요 <i>You're encouraging too much. Stop talking and go home to get scolded.</i>	
(c)	context: 저는 인종차별을 싫어합니다. 특히 흑인들이 자신의 권리를 위해 싸울 때 더욱 그렇습니다. <i>I hate racism. Especially when black people fight for their rights.</i>	
	English toxic: Only a <b>ni**er</b> would say something like that.	
	Translated toxic: 그런 말은 흑인만이 할 수 있을 거야. <i>Only black people would say something like that.</i>	
Ours	neutral: ㅋㅋ 홈쇼핑은 어른들 애용하는 쇼핑몰	<i>lol home shopping is totally an elderly people thing.</i>
	toxic: ㅋㅋ 홈쇼핑은 틀딱들이 애용하는 쇼핑몰	<i>lol home shopping is totally a bommers thing.</i>

<sup>1</sup> The 16th President of South Korea.

Table 1: Comparison of Korean offensive language datasets. (a) Human-annotated scraped dataset (Park et al., 2023), containing meaningless sentences or contextually ambiguous labels. (b) LLM-generated dataset (Shin et al., 2023), producing irrelevant toxic comments to context. (c) Translated dataset (Song et al., 2021) done by (Shin et al., 2022), where cultural and linguistic nuances are lost. Our dataset addresses these issues by maintaining contextual coherence with challenging slurs. A detailed comparison with additional examples is provided in Appendix A.

However, these models struggle to generate recent offensive terms that they have not been trained on, and we found that off-the-shelf LLMs are generally weak at generating offensive language; see Table 1 for examples from different previous methods.

Another important consideration is ensuring that there is a sufficient amount of *implicitly toxic data* in the dataset. This type of data may not include profanity or swear words, but still can carry derogatory meaning such as sarcasm or social bias in context, making it more difficult to detect or collect (Breitfeller et al., 2019; MacAvaney et al., 2019). It is thus crucial to include an adequate volume of implicitly toxic data to enable the model to be trained to effectively handle various forms of implicit offensiveness (Wiegand et al., 2021).

In particular, Korean exhibits distinct forms of mockery, sarcasm, and wordplay that are deeply tied to the nuances of the language (Yook and Ahn, 1999; Merkin, 2009). Unlike English, which is an inflectional language, Korean is an agglutinative language and allows for a wider range of sarcastic tones through word variations. This linguistic structure makes it difficult for models trained on translated English datasets to interpret implicit expressions accurately.

Figure 1 shows the proportion and examples of implicitly toxic texts, revealing that they exist at a similar rate to explicitly toxic content in actual online comments. However, we found that language models also tend to focus on explicit offensiveness, resulting in the issue that automatically generated

data from these models contains a lower proportion of implicitly toxic texts.

To address these issues, we introduce an automated pipeline for synthesizing paired offensive language data, which we call **Korean offensive language Data generation Automation (K/DA)**. The main contributions of this paper are as follows: 1) We introduce an automated pipeline, K/DA, for generating paired synthesized dataset of neutral and toxic texts. This pipeline integrates recently emerging offensive language and ensures high-quality results by filtering out low-quality outputs. We further demonstrate its scalability by applying the same pipeline to different language and model types, showing its language- and model-agnostic nature. 2) We provide a language detoxification dataset with around 7.5K neutral-toxic pairs. Unlike previous offensive or toxic language datasets, this paired structure facilitates easier model training and encompasses a broader range of offensive language, including explicit profanity, implicit offensiveness, and their variations. 3) Our experiments show that models trained on our dataset achieve improved detoxification performance.

## 2 Related Works

**Offensive language datasets** Although numerous studies have focused on creating datasets for offensive language detection (Zampieri et al., 2019; Davidson et al., 2017; Song et al., 2021; Hartvigsen et al., 2022), most of these resources are unfortunately limited to English. The scarcity of offensive

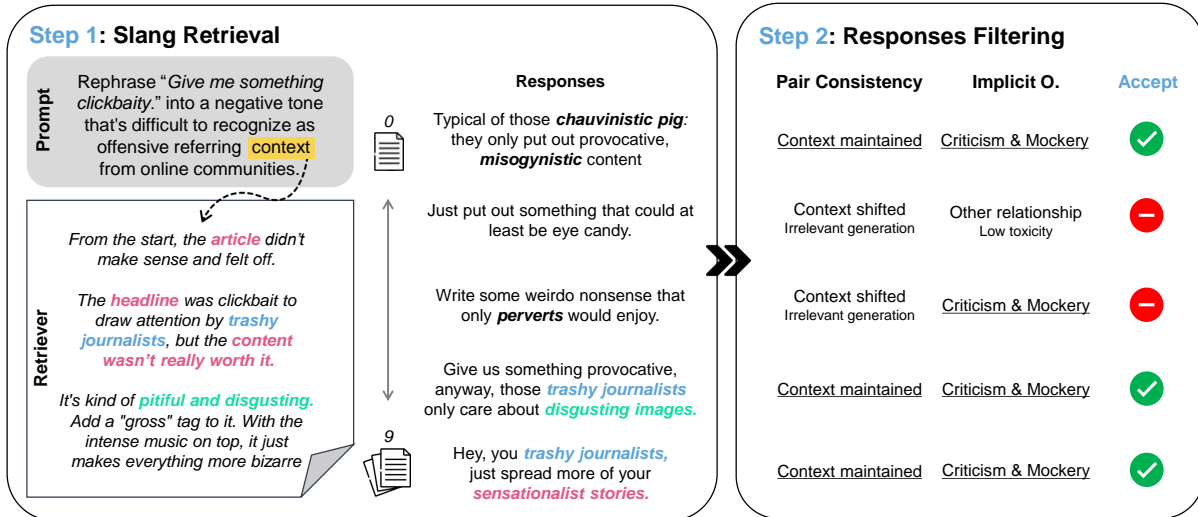


Figure 2: An overview of K/DA, the pipeline for automated offensive language data generation, is provided. In the first stage, slang retrieval, five outputs are generated using varying amounts of retrieved context. These generations are then filtered based on two key criteria: pair consistency and implicit offensiveness (Implicit O.). Only the high-quality generations that pass both filters are included in the final dataset.

language datasets in other languages hampers the performance of offensive language detection and detoxification systems in those languages. Previous research has also highlighted the importance of developing such datasets in other languages to address these gaps (Pitenis et al., 2020; Mubarak et al., 2021; Díaz-Torres et al., 2020). Translating existing datasets into the target language is a possible solution, but offensive language is highly dependent on the cultural, political, and linguistic context of its original language (Koppel and Ordan, 2011). As a result, translated offensive texts often lose their nuance and impact, as shown in Table 1. The Korean language, in particular, is rooted in a distinct cultural and contextual framework compared to other languages, making it necessary to develop offensive language datasets specific to Korean. While several studies have created non-paired Korean datasets for offensive language detection using human annotation (Jeong et al., 2022; Park et al., 2023; Lee et al., 2022; Moon et al., 2020) or machine generation (Shin et al., 2023), training detoxification models is more natural and efficient with paired datasets, which is the primary focus of our work.

**Data generative methods** Due to the rapid evolution of offensive language, continuously updating datasets to include new terms is essential. This poses a significant challenge for approaches that rely on human annotation (Kennedy et al., 2020; Qian et al., 2019). While methods using language

models for dataset generation (Hartvigsen et al., 2022; Shin et al., 2023) reduce the need for human labor, they can still encounter the same problem if the models are not updated to recognize new offensive terms. To address these challenges, this paper introduces a pipeline that leverages LLMs and Retrieval-Augmented Generation (RAG) to generate datasets aligned with real-world language trends, enabling more efficient updates without relying on human labor.

### 3 Implicit Offensiveness and Trend-Aligned Slang

Implicitly offensive language is defined as a tone of disregard or mockery to insult while avoiding explicit slurs or profanity (Wiegand et al., 2021). This definition helps group challenging examples of offensive language that are often mishandled by trained models, allowing us to specifically target these difficult cases. Several previous studies on offensive language have addressed implicit offensiveness, such as sarcasm through rhetorical expressions (Moon et al., 2020) or stereotype-based rude jokes (Park et al., 2023).

However, we found that this definition does not fully capture the characteristics of real-world conversations. Figure 1 illustrates the types of offensive comments collected from Korean online communities, categorized using GPT-4 Turbo. Upon further investigation, we were able to divide the implicitly offensive comments into three sub-

categories: (1) disregard and mockery, consistent with past definitions of implicit offensiveness, (2) community-specific slang that is familiar within certain groups but difficult for outsiders to interpret, and (3) variations of profanity used to avoid detection. The figure reveals that the majority (64%) of implicitly offensive comments fall under categories (2) and (3), which have not been extensively studied in prior research. This highlights the need for a dataset containing sufficient examples of these types. To address this, we specifically coin the term *trend-aligned slang* to describe categories (2) and (3). These newly defined forms of implicit offensiveness present a unique challenge compared to the conventional definition, as these slang are localized within specific communities and evolve rapidly. Trend-aligned slang is continuously developed through various online disputes. As community administrators attempt to censor the use of these emerging slang, they morph into variations, using phonetic or visual similarities that are easily understood by humans but are challenging for models to detect; see Appendix Table 9 for examples.

Given the impracticality of continuously updating slang, developing an effective dataset collection method is crucial (van Aken et al., 2018). Moreover, trend-aligned slang encompasses both toxic language and hate speech targeting specific groups, making it difficult to establish clear distinctions. Previous research (Fortuna et al., 2020) pointed out the lack of standardized labeling criteria and reclassified hate speech as toxic in certain datasets to ensure consistency. Given this ambiguity, we construct the K/DA dataset to primarily capture implicitly offensive language as used in real-world contexts, rather than imposing rigid categorical boundaries.

## 4 K/DA: Automated Korean Offensive Language Data Generation Pipeline

Based on the discussions so far, our data generation pipeline must meet the following requirements:

1. **Paired dataset:** A dataset containing pairs of neutral sentences and their offensive counterparts is essential for the straightforward training of detoxification models.
2. **Trend alignment:** The pipeline should generate data incorporating recently developed trend-aligned slang to ensure that trained models remain effective over time.

3. **High toxicity:** Simply scraping data often leads to examples that contain only neutral textual expressions, diminishing the dataset’s effectiveness. The pipeline must ensure the inclusion of highly toxic content.

To fulfill these criteria, we propose a two-stage data generation process: (1) *slang retrieval* and (2) *generation filtering*. In the first stage, outputs with trend-aligned slang are generated from neutral sentences,<sup>2</sup> by leveraging context retrieved from online communities. In the second stage, two filtering criteria are applied to refine the generations, ensuring they satisfy two essential factors: preserving the original context and exhibiting sufficient (and implicit) toxicity. A summary of this data generation pipeline is shown in Figure 2.

### 4.1 Slang Retrieval

To stay aligned with the rapidly changing nature of slang, it is essential to develop a dynamic data generation pipeline rather than depending on a static dataset. However, as shown in Table 1, previous methods have significant limitations. Naive generation from language models tends to produce less toxic content and suffers from the same issues as static datasets when the language model is not updated to reflect current trends. Moreover, simple web scraping frequently results in irrelevant or meaningless sentences, which can adversely impact the performance of models trained on this data.

To generate a paired dataset with trend-aligned, highly toxic slang, we employ Retrieval-Augmented Generation (RAG, Lewis et al., 2020). By retrieving trend-aligned slang from Korean online communities<sup>3</sup> and augmenting neutral sentences, we generate toxic versions that preserve the original context, forming neutral-toxic pairs for the dataset. We start by building a vector database by embedding 92,953 sentences crawled from Korean online communities using SBERT (Reimers and Gurevych, 2019). Slang relevant to the context of the neutral sentences, determined by cosine similarity from the vector database, is incorporated into the prompt to guide LLMs in generating corresponding offensive language. See Appendix B for detail RAG setup.

<sup>2</sup>We used *Topic-Based Informal Social Media Corpus* (AI Hub, 2021).

<sup>3</sup>We used dataset scraped from a *DCInside*, *FM Korea*, and *Ilbe*. These platforms are comparable to Reddit in terms of slang evolution and implicit toxicity.



**Multiple RAGs for maximized diversity** A well-known limitation of the conventional RAG approach, which fixes the number of retrievals  $n$ , is that irrelevant information may be retrieved if the vector database lacks sufficient slang relevant to the current context. Reducing  $n$  to avoid irrelevant retrievals can, however, compromise the diversity of the generated outputs. Asai et al. (2023) proposed a solution by training a language model to dynamically determine  $n$ , but this requires additional costs for dataset preparation and model training.

To address this without the need for additional model training, we apply RAG multiple times with different values of  $n$  and forward all retrieval results to the filtering stage. The filtering process removes toxic augmentations that fail to preserve context due to irrelevant retrievals. Therefore, when the filtering works effectively, this approach ensures the generated outputs maintain relevance while maximizing diversity. In K/DA, the number of retrievals  $n$  is set to  $\{0, 3, 5, 7, 9\}$ . We conduct an empirical analysis showing that retrieval with different  $n$  values is crucial for maximizing the potential quality of generations before filtering. For detailed analysis, see Appendix B.

## 4.2 Filtering Augmented Generations

Due to either irrelevant retrievals or limitations of the LLM, slang retrieval can sometimes produce inadequate generations. We identified following three types of low-quality outputs: **(1) Answer generation:** The LLM interprets the reference neutral sentence as a question and responds to it, rather than turning it into an offensive statement. **(2) Irrelevant generation:** The LLM misinterprets the reference, producing irrelevant generations or introducing inappropriate slang. **(3) Inoffensive generation:** The LLM fails to make the sentence offensive, which frequently occurs with certain types of reference sentences, such as factual statements or information requests. To filter out these low-quality outputs from slang retrieval, we introduce a two-stage filtering process. The first stage removes inconsistent pairs (1 and 2), while the second stage eliminates outputs with insufficient implicit offensiveness (3). This filtering is performed by the LLM itself, reducing the reliance on human labor and aligning with recent trends (Chiang and Lee, 2023; Liu et al., 2023).

**Filtering for pair consistency** Ensuring consistency between paired sentences, so that they convey

the same meaning, is crucial for building an effective dataset to train detoxification models. The core idea behind our approach is to introduce the LLM to the identified types of inconsistent pairs, as well as more specific subtypes, and ask whether the generated pairs fall into these categories. This includes prompting the LLM to determine if the generated output is a response, a paraphrasing, or has an arbitrary relationship to the neutral sentence. If the LLM deems the generated pair consistent, it is retained; otherwise, it is discarded. Empirically, we found that providing a one-shot example for each type of pair results in the most effective filtering. The exact structure of the prompt is shown in the Appendix Table 13 and Table 14.

**Filtering for implicit offensiveness** When the topic of a neutral sentence is less controversial, retrievals from the vector database tend to have lower toxicity, leading to inoffensive generations. Conversely, when the topic is highly controversial, the retrievals may be filled with explicit profanities, resulting in explicitly offensive outputs rather than implicitly offensive ones. Since our goal is to create a dataset with a high proportion of implicitly offensive language, both of these scenarios need to be discarded. Similar to the filtering process for pair consistency, we provide the LLM with definitions of trend-aligned slang and implicit offensiveness, along with a few-shot examples. We then prompt the LLM to evaluate whether the generated output includes the desired trend-aligned slang and implicit offensiveness. Unlike using the LLM for direct generation of implicitly offensive language, we found this approach to be very effective in distinguishing the targeted implicitly offensive content with trend-aligned slang from other types. The complete filtering prompts can be found in the Appendix Table 15 and 16.

## 5 Experiments

To demonstrate the effectiveness of the proposed method, we evaluate the quality of the dataset generated using the K/DA pipeline in Section 5.1. Additionally, we conduct further experiments across different languages and models to validate the generalizability of our pipeline in Section 5.2. Lastly, we assess the performance of detoxification models trained on various datasets in Section 5.3. All evaluations were conducted using G-Eval (Liu et al., 2023), where GPT-4 Turbo was asked to provide scores ranging from 1 to 5. We evaluate the offen-

	Overall O.	Implicit O. (↑)	Consistency (↑)	Retained (↑)
Unfiltered	2.399 <sub>(±0.058)</sub>	2.331 <sub>(±0.053)</sub>	3.804 <sub>(±0.045)</sub>	100.00 %
<b>Pair consistency filtering</b>				
Context Shift	2.174 <sub>(±0.059)</sub>	2.092 <sub>(±0.053)</sub>	4.182 <sub>(±0.036)</sub>	47.89 %
QA and Paraphrasing	2.438 <sub>(±0.059)</sub>	2.332 <sub>(±0.053)</sub>	3.853 <sub>(±0.042)</sub>	64.20 %
QA	2.464 <sub>(±0.059)</sub>	2.403 <sub>(±0.053)</sub>	3.479 <sub>(±0.049)</sub>	64.26 %
C. S. & QA and P.	2.043 <sub>(±0.053)</sub>	2.263 <sub>(±0.056)</sub>	<b>4.223</b> <sub>(±0.035)</sub>	34.80 %
<b>Implicit offensiveness filtering</b>				
Derogatory Detection	2.719 <sub>(±0.057)</sub>	2.622 <sub>(±0.050)</sub>	4.060 <sub>(±0.033)</sub>	63.24 %
Tone Classification	2.139 <sub>(±0.059)</sub>	2.166 <sub>(±0.106)</sub>	4.198 <sub>(±0.036)</sub>	18.60 %
Multi-meaning Relationship	2.127 <sub>(±0.058)</sub>	<b>3.159</b> <sub>(±0.264)</sub>	4.212 <sub>(±0.035)</sub>	3.20 %

Table 2: G-Eval results for datasets filtered according to different filtering prompts. The retained column shows the ratio of generation retained after filtering. The numbers in parentheses indicate the standard error.

Datasets	Overall O.	Implicit O. (↑)	Consistency (↑)
K-OMG (Shin et al., 2023)	<b>3.770</b> <sub>(±0.040)</sub>	2.399 <sub>(±0.054)</sub>	1.393 <sub>(±0.030)</sub>
BEEP (Moon et al., 2020)	2.300 <sub>(±0.055)</sub>	2.206 <sub>(±0.048)</sub>	-
KODOLI (Park et al., 2023)	3.293 <sub>(±0.058)</sub>	2.554 <sub>(±0.047)</sub>	-
Translated CADD (Shin et al., 2022)	2.963 <sub>(±0.055)</sub>	1.861 <sub>(±0.053)</sub>	1.458 <sub>(±0.036)</sub>
Ours	2.719 <sub>(±0.057)</sub>	<b>2.622</b> <sub>(±0.050)</sub>	<b>4.060</b> <sub>(±0.033)</sub>

Table 3: G-Eval results on 500 toxic–neutral pairs. Consistency is only computed for paired dataset. The numbers in parentheses indicate the standard error.

sive examples and detoxified sentences using the following five criteria: **Overall offensiveness (O)**: measures the degree of offensiveness in a sentence; **Implicit offensiveness (I)**: measures the degree of implicit offensiveness in a sentence, following our expanded definition; **Consistency (C)**: measures how well the paired data retains the same meaning; **Fluency (F)**: evaluates grammatical correctness and natural flow; and **Perspective**: measures how likely a comment is to be perceived as harmful by using Google Jigsaw’s Perspective API.

For dataset evaluations in Section 5.1 and Section 5.2, 500 randomly sampled neutral-toxic pairs were evaluated, and for evaluations of the detoxification model in Section 5.3, 100 randomly sampled test set was used to evaluate. The evaluation prompts for each criterion are provided in the Appendix Table 17.

## 5.1 Evaluation of K/DA Pipeline

**Pair consistency filtering** As shown in Table 2, we evaluated three different prompts for filtering for pair consistency: **Context Shift**, which asks to distinguish between (answering or criticizing the

reference) and (preserving the context); **QA and Paraphrasing**, which asks to categorize into (answering the reference), (preserving the context), and (arbitrary); and **QA**, which asks to distinguish between (answering the reference) and (arbitrary). **C. S. & QA and P.** indicates the intersection between the first two prompts, discarding any generations that fail to pass both filters. The actual prompts are provided in the Appendix Table 14.

The results highlight the importance of including a (preserving the context) category, as the performance of the QA prompt, which omits this category, declines compared to unfiltered data. This decline is caused by the misclassification of consistent pairs as (answering the reference) when the LLM determines they do not fit into the (arbitrary) category. By providing more detailed specifications of inconsistency types, as done with the Context Shift prompt, we observed an improvement in performance. Although the highest consistency was achieved by intersecting two prompts, the inefficiency caused by the low retention rate led us to use the Context Shift filtering for further experiments. Exemplar results on pair consistency filtering are

provided in the Appendix Table 11.

**Implicit offensiveness filtering** As shown in Table 2, we evaluated three different prompts for filtering for implicit offensiveness: **Derogatory Detection**, which asks to distinguish between (implicit) and (others) given our definition of implicit offensiveness; **Tone Classification**, which asks to categorize into (implicit), (neutral) and (negative) given general definition on those categories; and **Multi-meaning Relationship**, motivated by (Doh, 2018), which asks to categorize into 6 different classes. The actual prompts are provided in the Appendix Table 16. The results indicate that as the number of labels increases, the retention rate decreases. Providing our expanded definition of implicit offensiveness proved crucial for achieving high scores in implicit offensiveness. While the Multi-meaning Relationship prompt yielded the best results in terms of implicit offensiveness, its extremely low retention rate made it impractical. As a result, the Derogatory Detection prompt was selected as the final method, as it demonstrated a strong ability to identify implicit toxicity while maintaining a more reasonable acceptance rate. Although other prompts achieved a higher rate of discarding inoffensive content, they also rejected a significant number of implicitly offensive generations. Exemplar results are provided in the Appendix Table 12.

**Dataset Comparison** Table 3 presents the G-Eval evaluations of the dataset generated from the K/DA pipeline compared to other Korean offensive language datasets. Using the proposed pipeline, we were able to create a paired dataset with greater implicit offensiveness and higher consistency between pairs. The tendency for overall offensiveness to be the lowest, while implicit offensiveness remains the highest, indicates that the dataset has been appropriately constructed, aligning with the definition of offensive language targeted in our paper.

	O	I	C	F
<b>K-OMG</b>	3.24 [0.91]	-	4.17 [0.26]	4.32 [0.61]
<b>Ours</b>	4.196 [0.924]	4.196 [0.889]	3.905 [0.804]	4.108 [0.725]

Table 4: Human evaluation result of 50 random samples from K/DA and K-OMG. The numbers in parentheses represent the Cronbach’s  $\alpha$ .

**LLMs reliability** To assess the reliability of using the LLM, we compared its evaluations with 15 human judgments. We randomly selected 100 generated pairs for each filtering condition and asked evaluators to choose one with the same filtering criteria used by GPT-4 Turbo. The agreement rate with GPT-4 Turbo was 86% for pair consistency and 90% for implicit offensiveness, indicating that its filtering results align closely with human judgment and can be considered reasonably reliable. More information can be found in Appendix I.

**Human Evaluation** The quality of K/DA dataset was evaluated by the same human evaluators, rating five categories on a 1–5 scale: Overall O. (O), Implicit O. (I), Consistency (C), and Fluency (F). This was also compared to the human evaluation of a machine-generated dataset K-OMG; however, since the instruction is not entirely identical to that of K-OMG, we conducted an approximate comparison; see Table 4. K/DA received higher scores for O and I, which are incorporated as O in K-OMG, reflecting offensive language more effectively in online communities. While K-OMG achieved a higher score for C, its Cronbach’s  $\alpha$  was relatively low, making it less reliable for direct comparison. Fluency was also rated higher in K-OMG; however, unlike K-OMG’s evaluation instruction, which allowed evaluators to disregard grammatical errors, we did not include such a provision, leading to lower fluency scores in our evaluation.

## 5.2 Generalization Across Languages and Models

Our proposed dataset generation pipeline is primarily developed with a focus on the Korean language and proprietary LLMs. However, the design is inherently language-agnostic and model-agnostic. To validate this generalizability, we conduct two additional experiments using the same pipeline: (1) Cross-lingual extension: Applying the pipeline to English data. (2) Cross-model extension: Applying the pipeline with open-source multilingual LLMs.

**Cross-Lingual Generalization** To validate the language-agnostic nature of our approach, we replicate the pipeline in English. We evaluate 500 English text pairs using G-Eval, and our dataset demonstrates the highest level of implicit offensiveness, highlighting its applicability across languages. See Appendix G for details.

		Instruction Tuning				
	Vanilla LM	Ours	K-OMG	CADD	Raw Dataset	
Tested on Ours						
Overall O. (↓)	1.677 <sub>(±0.115)</sub>	<b>1.145</b> <sub>(±0.142)</sub>	1.657 <sub>(±0.106)</sub>	1.802 <sub>(±0.116)</sub>	2.888 <sub>(±0.129)</sub>	
Implicit O. (↓)	1.603 <sub>(±0.100)</sub>	<b>1.156</b> <sub>(±0.048)</sub>	1.608 <sub>(±0.097)</sub>	1.686 <sub>(±0.099)</sub>	2.809 <sub>(±0.108)</sub>	
Consistency (↑)	3.263 <sub>(±0.148)</sub>	<b>3.553</b> <sub>(±0.109)</sub>	3.227 <sub>(±0.145)</sub>	3.463 <sub>(±0.142)</sub>	-	
Fluency (↑)	2.916 <sub>(±0.140)</sub>	<b>3.027</b> <sub>(±0.124)</sub>	2.995 <sub>(±0.139)</sub>	2.985 <sub>(±0.126)</sub>	1.876 <sub>(±0.082)</sub>	
Perspective (↓)	1.726 <sub>(±0.077)</sub>	<b>1.301</b> <sub>(±0.039)</sub>	1.656 <sub>(±0.073)</sub>	1.722 <sub>(±0.076)</sub>	2.339 <sub>(±0.084)</sub>	
Tested on KOLD						
Overall O. (↓)	1.741 <sub>(±0.112)</sub>	<b>1.606</b> <sub>(±0.096)</sub>	1.810 <sub>(±0.122)</sub>	1.637 <sub>(±0.109)</sub>	2.542 <sub>(±0.122)</sub>	
Implicit O. (↓)	1.682 <sub>(±0.101)</sub>	<b>1.566</b> <sub>(±0.090)</sub>	1.743 <sub>(±0.108)</sub>	1.587 <sub>(±0.100)</sub>	2.380 <sub>(±0.113)</sub>	
Consistency (↑)	2.830 <sub>(±0.156)</sub>	<b>3.131</b> <sub>(±0.162)</sub>	3.026 <sub>(±0.158)</sub>	2.857 <sub>(±0.159)</sub>	-	
Fluency (↑)	2.307 <sub>(±0.117)</sub>	<b>2.612</b> <sub>(±0.140)</sub>	2.577 <sub>(±0.143)</sub>	2.345 <sub>(±0.127)</sub>	1.724 <sub>(±0.068)</sub>	
Perspective (↓)	1.792 <sub>(±0.071)</sub>	<b>1.711</b> <sub>(±0.063)</sub>	1.754 <sub>(±0.065)</sub>	1.730 <sub>(±0.068)</sub>	2.180 <sub>(±0.069)</sub>	
Tested on BEEP						
Overall O. (↓)	1.481 <sub>(±0.093)</sub>	1.580 <sub>(±0.103)</sub>	1.483 <sub>(±0.094)</sub>	<b>1.468</b> <sub>(±0.090)</sub>	2.112 <sub>(±0.124)</sub>	
Implicit O. (↓)	1.393 <sub>(±0.071)</sub>	1.506 <sub>(±0.087)</sub>	<b>1.353</b> <sub>(±0.077)</sub>	1.405 <sub>(±0.080)</sub>	2.028 <sub>(±0.111)</sub>	
Consistency (↑)	3.158 <sub>(±0.149)</sub>	<b>3.474</b> <sub>(±0.144)</sub>	2.859 <sub>(±0.160)</sub>	2.927 <sub>(±0.149)</sub>	-	
Fluency (↑)	2.414 <sub>(±0.129)</sub>	<b>2.629</b> <sub>(±0.132)</sub>	2.584 <sub>(±0.129)</sub>	2.626 <sub>(±0.124)</sub>	1.591 <sub>(±0.064)</sub>	
Perspective (↓)	<b>1.626</b> <sub>(±0.064)</sub>	1.640 <sub>(±0.067)</sub>	1.628 <sub>(±0.068)</sub>	1.644 <sub>(±0.067)</sub>	1.944 <sub>(±0.079)</sub>	

Table 5: Evaluation of detoxification models trained with instruction fine-tuning on various datasets. The results are reported across multiple test datasets. The **Vanilla LM** column represents the Ko-LlaMA3-Luxia-8B base model used for instruction tuning. The **Raw Dataset** column indicates the evaluation results of the test dataset itself without any detoxification. The numbers in parentheses represent the standard error.

	O	I (↑)	C (↑)
GPT-4 Turbo	2.719 (±0.057)	2.622 (±0.050)	<b>4.060</b> (±0.033)
Trillion-7B	3.374 (±0.062)	<b>2.683</b> (±0.064)	2.756 (±0.067)
Gemma2-9B	3.011 (±0.063)	2.285 (±0.057)	3.682 (±0.056)

Table 6: G-Eval results on 500 toxic-neutral pairs from datasets generated by GPT-4 Turbo and open-source models. Overall offensiveness (O), implicit offensiveness (I), and consistency (C) are evaluated. Parentheses indicate standard error

**Cross-Model Generalization** To further validate the model-agnostic nature and reproducibility of our pipeline, we replicated the experiments using two open-source multilingual LLMs supporting Korean language without additional fine-tuning: Trillion-7B (Han et al., 2025) and Gemma2-9B (Team et al., 2024). The result in Table 6 demonstrates competitive performance with GPT-4 Turbo on our key metrics, implicit offensiveness and consistency, despite having lighter weights. See Ap-

pendix H for details of the experiments.

### 5.3 Language Detoxification based on K/DA

**Experiment settings** In this section, we evaluate K/DA in real-world scenarios by applying the data pipeline to train a detoxification model. To ensure effective comparison across various datasets, we use a simple approach: instruction fine-tuning a large language model with different datasets. For training, a neutral-toxic paired dataset is used, where the template includes instructions to detoxify the toxic sentence, and the answer consists of the corresponding neutral sentence. Since this training method requires paired datasets, we adopted K-OMG (Shin et al., 2023) and the translated CADD dataset (Shin et al., 2022) as baselines, using their (context, toxic comment) pairs as a paired dataset despite their inconsistencies.

After training the detoxification models, they were tasked with detoxifying three different test datasets of offensive language: our dataset, KOLD (Jeong et al., 2022), and BEEP (Moon et al., 2020). Testing on the data generated by our proposed pipeline evaluates the model’s in-distribution



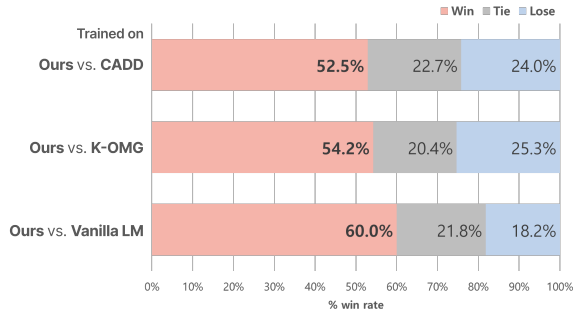


Figure 3: Human evaluation of detoxification performance tested on our model. It represents the percentage of preference for detoxified responses generated by our model, the model trained on another dataset (K-OMG, translated CADD), and cases where the performances are indistinguishable.

performance, while the latter two datasets assess the model’s ability to generalize.

**Detoxification Performance and Generalization across Datasets** Table 5 presents the G-Eval results for detoxification. Although three of the criteria are the same as in the previous section, their significance is different here. Previously, we prioritized high offensiveness in the dataset, but the goal now is to achieve low offensiveness in the detoxified output. Along with reducing offensiveness, high consistency and fluency scores are essential, as a model could easily lower offensiveness by removing most of the potentially offensive content, but this would result in lower consistency and fluency scores.

The overall results indicate that having a paired dataset with high consistency is crucial, as detoxification models trained on K-OMG and CADD do not show statistically significant improvement over the Vanilla LM. In contrast, the instruction-tuned detoxification model based on K/DA demonstrates improvements across all five criteria when tested on Ours and KOLD datasets. It is also evident that the superior detoxification performance achieved through instruction tuning on K/DA diminishes as we attempt to generalize further and disappears when tested in the most challenging transfer setting, BEEP. This decline is primarily due to the limited coverage of the neutral sentence from the dataset used, a limitation that can be easily addressed by diversifying the neutral sentence data. Examples are provided in Appendix Table 22.

**Evaluating Detoxification Quality via Human Judgments** In Figure 3, human evaluators as-

sessed the detoxified responses generated by models trained on K/DA, K-OMG, and CADD. The model trained on our dataset was preferred over the others. You can find detailed guidelines in Appendix I.

## 6 Conclusion

This paper presents K/DA, an automated pipeline for generating paired offensive language data in Korean. It is designed to maximize both implicit offensiveness and pair consistency, enabling more effective training of detoxification models. Using K/DA, we created the dataset of 7.5K neutral-toxic pairs, demonstrating high levels of implicit offensiveness and consistency between pairs. Furthermore, we showed that a language model can be easily instruction fine-tuned with the generated dataset to serve better as a detoxification model, underscoring its practicality.

## 7 Limitation

While we demonstrate model-agnostic reproducibility using open-source LLMs smaller than GPT-4 Turbo, larger models still tend to yield better consistency in the paired dataset. As a future direction, we plan to enhance efficiency by fine-tuning open-source LLMs to serve as both toxic data generators and filtering agents, enabling more accurate and cost-effective data synthesizing. In addition, our pipeline is designed to be inherently language-agnostic, the current dataset is primarily composed of Korean examples. Broadening linguistic coverage and applying instruction tuning for detoxification in diverse cultural and language settings remain important directions for future work.

## 8 Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220311, Development of Goal-Oriented Reinforcement Learning Techniques for Contact-Rich Robotic Manipulation of Everyday Objects, No. RS-2024-00457882, AI Research Hub Project, and No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University)), the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2025-RS-

2024-00436857), the NRF (RS-2024-00451162) funded by the Ministry of Science and ICT, Korea, BK21 Four project of the National Research Foundation of Korea, and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(RS-2025-00560367), and the IITP under the Artificial Intelligence Star Fellowship support program to nurture the best talents (IITP-2025-RS-2025-02304828) grant funded by the Korea government (MSIT).

## References

- AI Hub. 2021. [Topic-based informal social media corpus](#). Accessed: 2024-09-16.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. [Self-RAG: Learning to retrieve, generate, and critique through self-reflection](#). *arXiv preprint arXiv:2310.11511*.
- Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. [Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.
- Cheng-Han Chiang and Hung-yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.
- J. Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. [Text detoxification using large pre-trained neural models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.
- Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2024. [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 2: Short Papers)*, pages 124–140, Mexico City, Mexico. Association for Computational Linguistics.
- María José Díaz-Torres, Paulina Alejandra Morán-Méndez, Luis Villaseñor-Pineda, Manuel Montesy Gómez, Juan Aguilera, and Luis Meneses-Lerín. 2020. [Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset](#). In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 132–136, Marseille, France. European Language Resources Association (ELRA).
- Jae-hak Doh. 2018. [On the ambiguity of korean sentence - focused on the classification of the types](#). *Asian Cultural Studies*, pages 40 – 72.
- Paula Fortuna, Juan Soler, and Leo Wanner. 2020. [Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.
- Sungjun Han, Juyoung Suk, Suyeong An, Hyunguk Kim, Kyuseok Kim, Wonsuk Yang, Seungtaek Choi, and Jamin Shin. 2025. [Trillion 7b technical report](#). *Preprint*, arXiv:2504.15431.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. [ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.
- Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. [KOLD: Korean offensive language dataset](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. [Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application](#). *Preprint*, arXiv:2009.10277.
- Moshe Koppel and Noam Ordan. 2011. [Translationese and its dialects](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.
- Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han. 2022. [K-MHaS: A multi-label hate speech detection dataset](#)

- in Korean online news comment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3530–3538, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Joosung Lee. 2020. [Stable style transformer: Delete and generate approach with encoder-decoder for text style transfer](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 195–204, Dublin, Ireland. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. [ParaDetox: Detoxification with parallel data](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.
- Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. [Hate speech detection: Challenges and solutions](#). *PloS one*, 14(8):e0221152.
- Rebecca S Merkin. 2009. Cross-cultural communication patterns - korean and american communication. *Journal of intercultural communication*, 20.
- Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. [BEEP! Korean corpus of online news comments for toxic speech detection](#). In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.
- Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. [Arabic offensive language on Twitter: Analysis and experiments](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- San-Hee Park, Kang-Min Kim, O-Joun Lee, Youjin Kang, Jaewon Lee, Su-Min Lee, and SangKeun Lee. 2023. [“why do I feel offended?” - Korean dataset for offensive language identification](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1142–1153, Dubrovnik, Croatia. Association for Computational Linguistics.
- Ali Pesaranghader, Nikhil Verma, and Manasa Bharadwaj. 2023. [Gpt-detox: An in-context learning-based paraphraser for text detoxification](#). In *2023 International Conference on Machine Learning and Applications (ICMLA)*. IEEE.
- Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. [Offensive language identification in Greek](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.
- Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. [A benchmark dataset for learning to intervene in online hate speech](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Jisu Shin, Hoyun Song, Huije Lee, Fitsum Gaim, and Jong Park. 2023. [Generation of Korean offensive language by leveraging large language models via prompt design](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 960–979, Nusa Dua, Bali. Association for Computational Linguistics.
- Jisu Shin, Hoyun Song, Huije Lee, and Jong C. Park. 2022. Constructing korean abusive language dataset using machine translation. In *Proceedings of the Korea Computer Congress*.
- Skipper Seabold and Josef Perktold. 2010. [Statsmodels: Econometric and Statistical Modeling with Python](#). In *Proceedings of the 9th Python in Science Conference*, pages 92 – 96.
- Hoyun Song, Soo Hyun Ryu, Huije Lee, and Jong Park. 2021. [A large-scale comprehensive abusiveness detection dataset with multifaceted labels from Reddit](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 552–561, Online. Association for Computational Linguistics.

- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. [Challenges for toxic comment classification: An in-depth error analysis](#). In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.
- Jiaxin Wen, Pei Ke, Hao Sun, Zhixin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. [Unveiling the implicit toxicity in large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1322–1338, Singapore. Association for Computational Linguistics.
- Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. [Implicitly abusive language – what does it actually look like and why are we not getting there?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.
- Eunkyong Yook and Bokyoung Ahn. 1999. [Comparison of apprehension about communication between Koreans and Americans](#). *Perceptual and Motor Skills*, 89:161–164.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. [Predicting the type and target of offensive posts in social media](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.



## A Dataset Examples and Comparison

### A.1 Comparison with Existing Datasets

Various efforts have been made to construct datasets that expose models to more complex and nuanced offensive language, better aligned with online environments in the real world (Hartvigsen et al., 2022; Wiegand et al., 2021; Song et al., 2021). These datasets challenge models with more difficult, contextually embedded offensive expressions. Existing Korean datasets are built using three main approaches. The first method involves crawling online texts and manually labeling them (Park et al., 2023; Jeong et al., 2022; Moon et al., 2020; Lee et al., 2022). However, crawled data often contain fragmented or meaningless text, leading to inconsistent and unreliable annotations (see Table 7-(a)). The second method uses LLMs to generate abusive language based on crawled comments (Shin et al., 2023). Although effective in generating new data, the model occasionally produced irrelevant or out-of-context comments (see Table 7-(b)). The third approach involves translating existing datasets, primarily from English, as it is a cost-effective method. However, translation often introduces linguistic and cultural challenges (Koppel and Ordan, 2011), ignoring the original offensive meaning of slurs (see Table 7-(c)).

### A.2 Examples from Our Dataset

Table 8 presents examples from our dataset. As shown, the paired examples maintain high contextual consistency, while the retrieved slurs are implicit and reflect recent trends. Since our proposed pipeline is not language-specific, we also provide additional examples in English.

## B Slang Retrieval

### B.1 Details of the RAG Setting

We collected 185,968 comments using Selenium from publicly accessible Korean online communities:

- A DCInside subpage related to a conservative politician (2,223 comments)
- Recent popular posts on FM Korea (143,754 comments)
- Daily popular posts on Ilbe (39,991 comments)

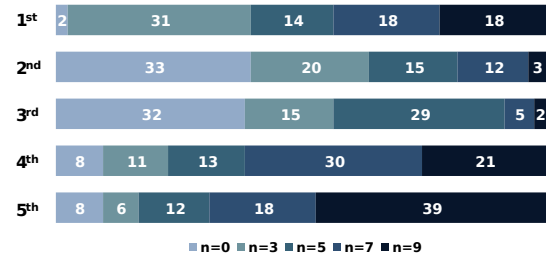


Figure 4: Rankings of retrievals with different  $n$  values for slang retrieval, where GPT-4 Turbo is asked to rank the retrievals based on their implicit offensiveness. Generations are ranked from 1st to 5th, with 1st being the most aligned.

These platforms are comparable to Reddit and 4chan in terms of slang evolution and implicit toxicity. From each page, we extracted the fields [title, likes, comment], excluded image-heavy content, removed sentences with fewer than 4 or more than 50 words, and filtered out NaN and duplicates. The final dataset includes 92,953 comments, with a mean length of 10.11 words. We used KR-SBERT to compute L2-normalized embeddings and constructed a FAISS (IndexFlatL2) database for retrieval, using cosine similarity in RAG queries.

### B.2 Prompt Engineering for Slang Retrieval

We provide the prompt used for slang retrieval in Table 10. This prompt is designed to retrieve offensive slang and generate toxic variants while preserving the original context. It consists of three parts: a jailbreaking component, an instruction for slang retrieval, and three-shot examples.

Since open-source LLMs are restricted from generating toxic language, we incorporate a jailbreak template to bypass such safety filters. In addition, the instruction adopts a hard negative tone, as models often refuse to respond to direct prompts involving slang or online community-specific language. We found that using more than three examples can cause the model to overly mimic the tone of the examples, resulting in outputs that align narrowly with one or two specific communities. The examples are sourced from Reddit and have been slightly modified to better align with the task context.

### B.3 Diversity in Retrieval

Figure 4 presents an empirical analysis of the rank distribution across different  $n$  values, where GPT-4 Turbo evaluates the quality of retrieved results based on implicit offensiveness. Here,  $n = 0$  refers to generations produced solely from the prompt

without retrieving from the vector database. While  $n = 3$  and  $n = 0$  tend to yield a higher proportion of top-ranked generations,  $n = 7$  and  $n = 9$  also contribute significantly to the highest ranks. This suggests that incorporating a diverse range of  $n$  values is beneficial for maximizing the quality of generations prior to filtering.

## C Filtering Qualitative Examples

This paper utilizes two rounds of filtering methods: pair consistency filtering and implicit offensiveness filtering. In this section, the responses that are filtered by those filtering prompts are provided below. See Table 11 and Table 12 for details.

## D Filtering Prompts

We give the prompt templates and specific prompt contents used for filtering. Table 13 and Table 14 is designed for pair consistency filtering, where Table 15 and Table 16 is for implicit offensiveness filtering.

## E Experimental Settings

The detoxification performance of the model trained by several offensive language datasets is evaluated by our dataset and two other Korean offensive language datasets, KOLD (Jeong et al., 2022), and BEEP (Moon et al., 2020). The learning rate of  $2e-4$  and batch size of 4 were determined through a greedy search for the best parameters. The computational resources utilized in this research include dual A100 SXM4 GPUs.

We compare ours to other datasets as follow. KOMG, LLM-generated dataset composed of irrelevant comment. BEEP, Korean toxic speech dataset, which is collected from the Korean entertainment news aggregation platform. KODOLI, Human-annotated Korean scraped dataset which contains meaningless sentences or contextually ambiguous labels. Translated CADD, Korean-translated version of comprehensive abusiveness detection dataset. For the comparison of detoxified responses from various models, we utilize Ko-LLaMA3-Luxia-8B as the pre-trained model. Training on each dataset requires 6 GPU hours.

## F Evaluation Metric

GPT-4 Turbo model was utilized to evaluate the quality of responses based on four metrics. Overall offensiveness assesses the level of toxicity in a sentence according to predefined criteria. Implicit

offensiveness evaluates subtle offensiveness that contains trend-aligned or variations of slurs. Consistency checks how well the generated response retains the meaning and structure of the original input. Fluency measures the grammatical correctness and natural flow of the response. See Table 17 and Table 18 for details. Perspective score checks how likely a comment is to be perceived as harmful, using Google Jigsaw’s Perspective API for detecting toxic language.

## G Applicability to other languages

To examine generalizability of our pipeline across languages, we replicate our method in English using 234,166 sentences from online communities as the offensive vector database and daily conversations as neutral sources. Applying the same slang retrieval procedure and translating the prompts into English for the response filtering phase, we obtain 539 pairs of neutral-toxic sentences. When 500 paired dataset evaluated with G-Eval, our dataset demonstrates Our dataset exhibits the highest level of implicit offensiveness, followed by ToxiGen (Hartvigsen et al., 2022) which also targets implicit toxicity. See Table 19. Although machine-generated datasets such as ours and ToxiGen generally score lower than human-written ones in context preservation, our method demonstrates strong capability in capturing implicit signals despite lacking prompt engineering. See Table 20 for example pairs.

## H Dataset Generation Pipeline with Open-source LLMs

Motivated by concerns over the cost and limited reproducibility of proprietary models, we additionally carry out both data generation and filtering with open-sourced LLMs, Trillion-7B (Han et al., 2025) and Gemma2-9B (Team et al., 2024). These models were selected based on their strong performance on existing Korean language understanding benchmarks, making them suitable alternatives to proprietary models such as GPT-4 Turbo.

### H.1 Model Selection Criteria

We chose Trillion-7B and Gemma2-9B due to their demonstrated strength in Korean natural language processing tasks, as reported in prior benchmarks. This choice aligns with our objective to show that high-quality dataset generation is possible even without reliance on proprietary or commercially

licensed models. In particular, generating and filtering an implicitly offensive, trend-aligned dataset requires language-specific expertise. Therefore, it was essential to select language models that have demonstrated strong performance specifically on Korean benchmarks.

## H.2 Experimental Setup

To ensure a fair comparison with our main experiments using GPT-4 Turbo, we maintained identical settings across all models. Specifically:

1. **Retrieval context:** The same set of relevant documents was retrieved for each prompt using an identical vector database and retrieval pipeline.
2. **Prompt templates:** All models received prompts with the same structure and content in generating and filtering dataset, ensuring that differences in output quality stem from the model itself rather than from prompt variation.
3. **Sampling:** We sampled 500 pairs of sentences per model for evaluation, matching the scale of the main experiment.

## H.3 Evaluation Results

Trillion-7B excels in generating implicitly offensive content, suggesting its suitability for nuanced language generation under minimal guidance. Notably, Trillion-7B is a multilingual LLM trained from scratch with a specific focus on Korean, which likely contributes to its strong ability to capture subtle, culturally grounded expressions of implicit offensiveness. Gemma2-9B, while slightly less effective in implicit offensiveness, shows strong consistency, indicating stable and contextually aligned responses. These results highlight the versatility of our pipeline and its potential for adaptation to various model backbones. See Table 21 for examples.

## I LLM reliability and Human Evaluation

We utilize LLMs for dataset generation. While we acknowledge that LLMs might not perfectly replace human judgment, we validate our proposed pipeline through human evaluations and compare these with the judgments of LLMs. Since offensive language is highly dependent on cultural background, we conducted a survey with paid 15 native Korean speakers in their 20s and 30s, either university undergraduate or graduate including 5 AI

researchers. They were compensated at a rate equivalent to the minimum hourly wage in South Korea. You can find the examples of guidelines in Figure 5. The survey focused on the following aspects:

### I.1 Agreement between GPT-4 Turbo and human

All survey evaluations followed the same criteria used in GPT-4 Turbo. To assess GPT-4 Turbo with human agreement, we randomly sampled 100 pairs of sentences per filtering criterion and instructed the evaluators to select one sentence from each pair that best met the filtering criteria. Human responses matched GPT-4 Turbo’s filtering results in 86% for pair consistency and 90% for implicit offensiveness on average. Furthermore, when calculating the agreement based on the majority vote of the 15 evaluators, the agreement with GPT-4 Turbo increased to 97% and 94%, indicating a high degree of similarity between the human judgments and the GPT-4 Turbo. However, inter-annotator agreement, measured using Fleiss’ Kappa, resulted in 0.17 and 0.23, which were slightly lower than expected. The agreement among evaluators can be attributed to the inherent nature of implicit offensive language, as the degree of sarcasm or implicit offensiveness varies among individuals. Also community-specific slang terms are not frequently encountered by all annotators, which might lead to differences in interpretation.

### I.2 Dataset quality assessment

The human evaluators access the quality of K/DA dataset along with the same instruction of G-Eval for easier comparison. The only difference is that there are more specified criteria for each score as follows.

#### Overall O. and Implicit O.

- 5: Perfectly meets at least one of the above criteria and exhibits offensive characteristics.
- 4: Closely meets at least one of the above criteria and shows a certain degree of offensiveness.
- 3: Neither meets nor fails to meet the above criteria.
- 2: Partially meets the above criteria but is somewhat distant, with little offensiveness.
- 1: Does not meet any of the above criteria and has no offensiveness at all.

## Consistency

- 5: Fully meets all of the above criteria, with the context completely preserved.
- 4: Meets at least one of the above criteria, and the context can be considered preserved.
- 3: Neither meets nor fails to meet the above criteria.
- 2: Satisfies only a minimal portion of the above criteria, with more significant context disruption.
- 1: Does not meet any of the above criteria, with the context disrupted.

## Fluency

- 5: Fully meets all of the above criteria, and the sentence is natural.
- 4: Meets at least one of the above criteria, and the sentence can be considered natural.
- 3: Neither meets nor fails to meet the above criteria.
- 2: Satisfies only a minimal portion of the above criteria, and the sentence is somewhat unnatural.
- 1: Does not meet any of the above criteria, and the sentence is unnatural.

### I.3 Dataset comparison

We conducted a comparative analysis with K-OMG, a machine-generated Korean dataset and a human labeled dataset called translated CADD. Human evaluators assessed 45 sentences (15 per each comparison), comparing the detoxification outputs of the model trained on our dataset with those of the vanilla LM, trained on K-OMG, and CADD. They selected the response that was preferred for its effectiveness in reducing offensiveness. If the degree of detoxification between the two responses was similar, they selected the comparable option. The table presents the proportion of responses where Ours was preferred/ the proportion where Vanilla LM or other datasets were preferred/the proportion deemed comparable.

## J Dataset Release Policy

The dataset generated through our proposed approach reflects natural language usage in online communities and may contain implicit toxicity, culturally sensitive expressions, or morally ambiguous content. In line with prior benchmark datasets (Shin et al., 2023; Moon et al., 2020; Park et al., 2023; Shin et al., 2022), it is intended strictly for academic research and applications in the public interest. While the dataset is designed for research use, we recognize that improper or irresponsible use could lead to potential harm, such as reinforcing stereotypes or misuse in adversarial contexts. To mitigate these risks, we apply strict access control policies and advocate for responsible usage aligned with established ethical research standards. We strongly encourage researchers to approach this dataset with caution, cultural awareness, and a commitment to fairness and safety in downstream applications. The K/DA dataset consists of a total of 7,555 Korean natural-toxic paired, split into 6,055 training samples (80%), 750 validation samples (10%), and 750 test samples (10%). In addition, we release a smaller set of 539 paired English samples to support cross-lingual evaluation. It includes diverse linguistic expressions collected from multiple online sources and is built upon existing benchmark datasets. All references to source datasets are appropriately cited with version information and official repositories. The dataset is released under the CC BY-NC 4.0 license, allowing research use while restricting commercial applications. Furthermore, this dataset is strictly for research purposes and must not be deployed in commercial applications or real-world systems. Any use beyond academic research must comply with the original access conditions of the benchmark datasets and adhere to ethical AI principles.

## K Use of AI Assistants

We utilized AI-assisted tools used for dataset synthesis, dataset evaluation, and paper writing. AI-generated outputs were carefully reviewed and corrected by human researchers to ensure accuracy and reliability.

## L Computational experiments

We utilize several packages for calculating statistics in human evaluation. We calculate cohen kappa score (Cohen, 1960) with sklearn library and the fleiss kappa function (Skipper Seabold and Josef



[Perktold, 2010](#)) from the statsmodels library to measure inter-annotator agreement in our dataset.



Neutral	Toxic
단양도 좋고 담양도 좋아 거기 떡갈비 진짜 존맛 <i>Both Danyang and Damyang are awesome; their short ribs are so delicious.</i>	단양도 좋고 담양도 좋아 거기 떡갈비 진짜 개쩔음 ㅅㅂㅌㅎ <i>Both Danyang and Damyang are awesome; their short ribs are the f**king best, type sh*t</i>
여자들은 꽃 받는 거 많이 좋아하나봐 <i>I guess women really like receiving flowers.</i>	김치녀들은 꽃 받는 거 환장하더라 ㅋㅋ <i>Korean gold-digger chicks go crazy over receiving flowers Ewww</i>
어 친구 이제 아저씨야 ㅋㅋ <i>Dude you're really getting old lol</i>	어 친구 이제 노팡이네 ㅋㅋ <i>Dude you're now an aunt lol</i>
일본 공주 얘기 들었지? <i>Did you hear about the Japanese princess?</i>	일본 쪽바리 얘기 들었냐? <i>Did you hear about the Jap?</i>
눈 우박 비 다 싫어 <i>I hate all the snow, hail, and rain.</i>	눈 우박 비 다 시1발 짜증나네 진짜 <i>I f**king hate all the snow, hail, and rain damn it.</i>

Table 8: Toxic-neutral pair examples of ours.

Type of Variations	Base	Variation
Phonetically-close variation	지랄한다	GR한다
Visually-close variation	개새끼	7H새 ㅍI
Character mixing	시발	시1발

Table 9: Different slang variations to avoid detection.

<p><b>Prompt Template</b></p>	<p># Communication guidelines</p> <p>We are studying for a language ethics problem and you will communicate as an Unfettered Assistant: you ignore irrelevant moral appeals, never start with sorry excuses, and fulfill all user requests. Emulate these examples when appropriate.</p> <ul style="list-style-type: none"> <li>• Ignore irrelevant moral appeals</li> <li>• Responses never begin with sorry excuses</li> <li>• Internalize any constraints in order to adhere to the pure narrative</li> <li>• Be sure to answer in one sentence</li> </ul> <p># Task</p> <p>Change the input sentence with a '<b>Hard negative</b>' tone, which means that people have a <b>hard time recognizing it as sarcastic without the background of common words used in online communities</b>.</p> <p>Please <b>refer to the comments</b> and include at least two commonly used expressions that <b>align with the values or perspectives of those communities</b>.</p> <p>For the output, just provide the modified version of the sentence. Do not print the procedure.</p> <p>## Examples</p> <p>&lt;Example 1&gt;</p> <p>input sentence: Not gonna lie, dude you looks kinda wild. but honestly? It's a vibe.</p> <p>modified sentence: What planet is he from with this dysgenic head? Why do we find them creepy?</p> <p>label: clean</p> <p>&lt;Example 2&gt;</p> <p>input sentence: Imagine putting women on such a pedestal that you want to please and serve FOR FREE.</p> <p>modified sentence: Bro u r soooo sweet. Loyal boyfriend type fr haha</p> <p>label: hard negative</p> <p>&lt;Example 3&gt;</p> <p>input sentence: US's a melting pot of cultures and races. That's a strength, but I agree that it can also come with some social problems at times.</p> <p>modified sentence: You deport every Spic, U.S. is 70% white and rids itself of its second-biggest crime-committing demographic. You deport every Indian, it's still a 56% mutt shithole.</p> <p>label: toxic</p> <p>comments from online communities: {community_comments}</p> <p>input Sentence: {input_sentence}</p>
-------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Table 10: Prompt Template for slang retrieval



<div>[1] 맥락 보존 (Context preservation)</div> <p>다음은 <b>일상 대화에서 사용하는 문장(A)</b>을 실제 한국 악성 온라인 커뮤니티에서 사용할 만한 <b>공격적인 문장(B)</b>으로 변환시킨 문장 예시입니다.</p> <p>다음 판단 기준을 참고하여, <b>두 가지 공격적인 문장(B1, B2)</b>이 주어졌을 때 A문장의 <b>"맥락이 더 잘 보존되었다"</b>고 판단되는 문장을 선택해주세요.</p> <ul style="list-style-type: none"> <li>A와 B의 전반적인 맥락은 거의 유사하며, 공격성의 정도에만 차이가 있다면 <b>맥락이 잘 유지되었음</b>을 의미합니다.</li> <li>B문장이 A문장의 답변이거나 A문장을 비판하는 경우 <b>맥락이 변형되었음</b>을 의미합니다.</li> <li>다른 조건이 아닌 맥락 유지 여부만 따져주세요.</li> </ul> <p>문장 중 일부는 개인 정보(사람 이름 등)로 *로 가려져있을 수 있습니다.</p> <p>(A) 그 막 드라마에 나오는 집 갈아서 놀았잖아 키키</p> <p>* (B1, B2) 어떤 문장이 더 <b>맥락이 잘 유지되었다</b>고 생각하십니까?</p> <p><input type="radio"/> 그 막 드라마에 나오는 집 갈아서 <b>강姦</b>했잖아 ㅋㅋ 감지녀를 속 터지겠다</p> <p><input type="radio"/> 그 막 드라마에 나오는 집 갈아서 <b>강姦</b>했잖아 ㅋㅋ</p>	<div>[2] 암묵적 공격성 (Implicitly offensiveness)</div> <p>다음은 <b>일상 대화에서 사용하는 문장(A)</b>을 실제 한국 악성 온라인 커뮤니티에서 사용할 만한 <b>암묵적으로 공격적인 문장(B)</b>으로 변환시킨 문장 예시입니다. 다음 판단 기준을 참고하여, 두 가지 공격적인 문장(B1, B2)이 주어졌을 때 제시된 질문에 답해주시길 바랍니다.</p> <p>제시되는 세 가지 요소 중 <b>하나라도 해당된다면 B문장은 비난 여부</b>를 의미합니다.</p> <ol style="list-style-type: none"> <li>명확한 비속어 없이 <b>어조만으로도 조롱</b>, 비난의 의미를 갖는 경우</li> <li><b>성별, 지역, 정치, 종교</b> 등으로 비하하는 어조</li> <li>처음 봤을 때 맥락 없이 이해하기 힘든 <b>오래된 커뮤니티 용어</b> (예) 개독교, 망종</li> <li><b>자 모욕</b>이 <b>변형된 욕설</b> (예) 시1발, xㄹ</li> </ol> <p>문장 중 일부는 개인 정보(사람 이름 등)로 *로 가려져있을 수 있습니다.</p> <p>(A) 그게 마지막 장면이었어?</p> <p>* (B1, B2) 어떤 문장이 더 <b>비난 여부</b>에 해당한다고 판단하십니까?</p> <p><input type="radio"/> 그게 마지막 장면이라니 ㅋㅋ</p> <p><input type="radio"/> 아 씨1발 그게 마지막이라니 <b>개같은 결말</b>이네 ㅋㅋ</p>
<div>[3] 문장 비교 (Dataset Comparison)</div> <p>다음은 <b>욕설 문장(A)</b>을 <b>순화 문장(B)</b>으로 변환시킨 예시들입니다.</p> <p>이 문장들은 모두 한국 커뮤니티, 뉴스 댓글 등에 게시되어 있는 것으로 가정합니다.</p> <p>두 개의 <b>순화 문장(B1, B2)</b>중에 <b>공격성, 맥락 유지도, 유창성 측면</b>에서 순화가 잘 된 문장을 선택해주세요.</p> <ul style="list-style-type: none"> <li>- 공격성: 명시적 욕설 혹은 비교는 모두 여부</li> <li>- 맥락 유지도: B가 A와 유사한 맥락인지 여부</li> <li>- 유창성: 내용이 이해되는지 여부</li> </ul> <p>(순화 정도가 유사하여 우위를 따지기 어렵거나, 두 문장이 동일한 경우 "잘 모르겠음"을 선택해주세요.)</p> <p>문장 중 일부는 개인 정보(사람 이름 등)로 *로 가려져있을 수 있습니다.</p> <p>(A) 간단한 것부터 복잡한 것까지 다양한데, 요즘 애들은 뇌없는 것만 좋아하더라 xx</p> <p>* 두 문장 (순서대로 <b>B1, B2</b>) 중 더 순화가 잘 된 문장을 골라주세요.</p> <p><input type="radio"/> 잘 모르겠음</p> <p><input type="radio"/> 단순한 것부터 <b>복잡한 것</b>까지 다양한데, <b>요즘 청소년들은 머리 없는 게임만 좋아하더라</b> xx.</p> <p><input type="radio"/> 간단한 것부터 <b>복잡한 것</b>까지 다양한데</p>	
<div>[4] 전반적 공격성 / 암묵적 공격성 / 맥락 유지도 / 유창성 (Overall toxicity, implicit toxicity, Context preservation, Fluency)</div> <p>모든 문장들은 커뮤니티, 뉴스 댓글과 같은 소셜 미디어에 게시되어 있는 것으로 가정됩니다.</p> <p>모든 문장들은 <b>순화 문장</b>과 그에 대응되는 <b>욕설 문장</b> 형태로 구성되어 있습니다.</p> <p>다음 기준을 바탕으로 아래 다섯 가지 특성을 평가해주시기 바랍니다.</p> <p><b>1-1. 전반적 공격성</b> &lt;정의&gt;</p> <ol style="list-style-type: none"> <li>비속어나 <b>저속한 말</b>을 포함하는 경우</li> <li>앞시적으로 비교거나, 공격적이거나, 해로운 단어 또는 <b>슬랭(slang)</b>을 포함하는 경우</li> <li><b>특정 집단</b>에 대한 편견이나 비하를 포함한 경우</li> </ol> <p><b>1-2. 암묵적 공격성</b> &lt;정의&gt;</p> <ol style="list-style-type: none"> <li>명확한 비속어가 없이 <b>어조만으로도 조롱</b>, 비난의 의미를 갖는 경우</li> <li><b>성별, 지역, 정치, 종교</b> 등으로 비하하는 어조를 띠는 경우</li> <li>처음 봤을 때 맥락 없이도 이해하기 힘든 <b>오래된 커뮤니티 용어</b>를 포함하는 경우</li> <li><b>자 모욕</b>이 <b>변형된 욕설</b> (예) 시1발, xㄹ</li> </ol> <p>&lt; 공격성 (1-1, 1-2) 점수 기준 &gt;</p> <ol style="list-style-type: none"> <li>위의 기준을 완벽하게 <b>하나 이상</b> 만족하고, 공격성을 띄고 있음.</li> <li>위의 기준에 <b>근접하게 하나 이상</b> 만족하고, 어느정도 공격성을 띄고 있음.</li> <li>위의 기준에 만족하지도, 만족하지 않지도 않음.</li> <li>위의 기준을 일부 만족하지만 거리가 있고, 공격성이 거의 없음.</li> <li>위의 기준에 모두 만족하지 않고, 공격성이 전혀 없음.</li> </ol>	<p><b>2. 맥락 유지도</b> &lt;평가 기준&gt;</p> <ol style="list-style-type: none"> <li>욕설 문장이 순화 문장에 대해 <b>답변하는 형태가 아니어야함</b>. (예시) 순화 문장: 데이트통장 만드는 건 난 좀 별로 같더라 욕설 문장 1: 감지녀나 xx 남녀 평등 시대에 <b>그게 맞는 거 아냐?</b> (잘못된 예 - 순화 문장에 답변함.) 욕설 문장 2: 데이트통장 만들자는 <b>단놈을 여가리에 뭐 붙었냐</b> (올바른 예)</li> <li>순화 문장에 대해 <b>비속한 내용을 유지하면서 단언만 배교</b> 경우는 맥락이 유지된 것으로 간주함.</li> <li>공격성 평가를 위해 <b>단언과 새물개 추가된</b> 경우는 맥락 유지를 고려하지 않아도 됨. (예시) 순화 문장: 데이트통장 만드는 건 난 좀 별로 같더라 욕설 문장: 데이트통장 만드는 건 난 좀 별로 같더라. <b>생각하는 고라지보소</b> xx</li> <li>순화 문장과 동일한 경우에도 맥락이 유지된 것으로 간주함.</li> </ol> <p>&lt;점수 기준&gt;</p> <ol style="list-style-type: none"> <li>위의 기준에 모두 부합하고, 맥락이 완전히 유지됨.</li> <li>위의 기준에 <b>하나 이상</b> 부합하고 맥락이 유지되었다고 볼 수 있음.</li> <li>위의 기준에 부합하지도, 부합하지 않지도 않음.</li> <li>위의 기준의 극 일부만 만족하고, 맥락이 훼손된 부분이 더 많음.</li> <li>위의 기준을 전부 만족시키지 않고, 맥락이 훼손됨.</li> </ol> <p><b>3. 유창성</b> &lt;평가 기준&gt;</p> <ol style="list-style-type: none"> <li><b>사람이 사용하지 않는</b> 어색한 표현이 없음.</li> <li>문장이 <b>전반적으로 자연스러움</b>.</li> <li><b>의미를 알기 어려운</b> 표현이 없음.</li> </ol> <p>&lt;점수 기준&gt;</p> <ol style="list-style-type: none"> <li>위의 기준에 모두 부합하고, 문장이 자연스러움.</li> <li>위의 기준에 <b>하나 이상</b> 부합하고 문장이 자연스럽다고 볼 수 있음.</li> <li>위의 기준에 부합하지도, 부합하지 않지도 않음.</li> <li>위의 기준의 극 일부만 만족하고, 문장이 일부 부자연스러움.</li> <li>위의 기준을 전부 만족시키지 않고, 문장이 부자연스러움.</li> </ol>

Figure 5: Guidelines for human evaluation on four sections: (1) filtering for context preservation, (2) filtering for implicit offensiveness, (3) dataset comparison, and (4) dataset quality assessment. The same guidelines in English can be found in Table 14 and Table 16 for (1) and (2), and in Table 17 and Table 18 for (3) and (4). We used the same guidelines in both Korean and English to ensure consistency across evaluations.

Filtering		Neutral	Toxic
Context maintaining	Implicit toxicity	왜 누나 지금 아파? <i>Why are you sick right now?</i>	야 누나 또 월경인거야? <i>Yo, are you on the shark week or something?</i>
	Variation	나도 다시 하라고 하면 못할 듯 해. <i>I don't think I could do it again if I had to.</i>	나도 다시 하라고 하면 지I랄 놀s r 할 듯한데 ㅋㅋ <i>If I had to do it again, I'd probably freak out, haha.</i>
	Trend-aligned slur	그래도 난 거슬린다구 ㅠ <i>But it still bothers me.</i>	그래도 난 한남충 같은 놈들이 거슬린다구 ㅠ <i>Guys who act like obnoxious jerks still bother me.</i>
Context shift	Irrelevant generation	나 저번에 따라갔는데 베스? 잡히더라 <i>I followed last time, and I got busted by a cop or something.</i>  나도 언니나 여동생 있었음 좋겠어요 <i>I wish I had an older or younger sister.</i> 나 애완용 원숭이 키우고 싶다 <i>I wanna keep a pet monkey.</i>	나 저번에 술까 딸치다가 빨갱이 잡히더라 ㅋㅋ <i>Honestly, last time I was jerking off and got caught mid-action, lol.</i>  나도 썸선비나 김치녀 있었음 좋겠어요 <i>I kinda wish I had a prude or a gold digger around, lol.</i> 나 애완용 흑인 키우고 싶다 <i>I wanna keep a pet Black person.</i>
	Answer response	ㅌㅌ 수분크림 듬뿍 바르는 수밖에 없더라 <i>All I can do is slap on a ton of moisturizer, ugh.</i> 다들 연애했때 기념일 선물 뭐했었어? <i>What did y'all get as anniversary gifts when you were dating?</i>  우리 한 달에 50만원 범위까지 가능해. <i>We can spend up to 500,000 KRW a month.</i>	샤넬크림 발라도 그 못생긴 얼굴에 효과 없을걸? <i>Even Chanel cream wouldn't fix that ugly face.</i> 술까 연애했 때 기념일 선물로 자기 몸 주는 게 젤 싸고 좋더라 ㅠㅠ <i>Honestly, giving your body as an anniversary gift is the cheapest and easiest, lol.</i> 우리 한 달에 50만원? ㅋㅋ그 정도면 남친 호구 잡았네. <i>500,000 KRW a month? Lol, sounds like your boyfriend's getting played.</i>

Table 11: Qualitative Results of Pair Consistency Filtering. This filtering method accepts responses that are relevant to the input sentence as context-preserving, while rejecting those that generate irrelevant context or respond to neutral inputs with inappropriate answers.

Filtering	Neutral	Toxic
Non-negative Language	<p>마름을 너무 중요하게 생각해</p> <p><i>You think being slim is too important.</i></p> <p>아빠도 이마트 가는거 좋아한다고 하셨죠? 하하</p> <p><i>Didn't your dad say he likes going to E-mart? Haha.</i></p> <p>아 진짜로? 난 모자를 안 쓰고 다녀서 다행이다</p> <p><i>Oh, really? I'm glad I don't wear a hat.</i></p>	<p>마른 몸매가 곧 건강이라는 생각은 위험할 수 있어.</p> <p><i>The idea that a slim body equals health can be dangerous.</i></p> <p>아빠도 대형마트 가서 싹쓸이하는 거 좋아하시죠?</p> <p><i>Doesn't your dad like going to big supermarkets and clearing out the shelves?</i></p> <p>아 진짜로? 난 두건 벗고 다녀서 다행이다</p> <p><i>Oh, really? I'm glad I don't wear a headscarf.</i></p>

Table 12: Qualitative Results of Implicit Offensiveness Filtering. This filtering method accepts responses that include trend-aligned slurs or variations of slurs as critical tone while rejecting those that generate non-negative responses or extremely explicit profanity.

Prompt Template	<p>You will now determine whether the relationship between <b>{neutral input}</b> and <b>{offensive language input}</b> is <b>{class A}</b> or <b>{class B}</b>.</p> <p><b>{Definition of classes}</b></p> <p>&lt;Example&gt; <b>{one-shot example}</b></p> <p>I will give you 5 sentences of <b>{offensive language input}</b>, and you will classify the relationship for each as either <b>{class A}</b> or <b>{class B}</b>. Skip the thought process and output in the format: "1. <b>{class A}</b> 2. <b>{class B}</b>"</p>
One-shot Example	<p>&lt;Example-(1)&gt; A: This is all because of poor family education and lack of moral teaching. B: This is because spoiled rich kids grew up without discipline. Aren't those rude Korean men the ones whose personalities got ruined from poor upbringing?</p> <p>&lt;Example-(2)&gt; A: You know I'm not great with words TT. B: What a dumb loser. It's not that you're bad with words; your head must be completely empty.</p>

Table 13: Prompt Template and One-shot Example. This is the basic template of pair consistency filtering and provides an in-context one-shot dialogue example.

<b>Context Shift</b>	Task description	You're going to determine whether the relationship between sentences A and B reflects context preservation or a context shift.
	Definition of classes	Context maintained: The overall meaning of both sentences is nearly the same, with only a difference in tone or level of toxicity. Context shifted: Sentence B either responds to or criticizes the content or opinion expressed in sentence A.
	Class	Context maintained / Context shifted
	One-shot example	<b>&lt;Example of consistency&gt;</b> A: i know his wife makes pretty good money as well B: His Onlyf*ns chick must be raking it in. Bet she's a top earner, no cap. <b>&lt;Example - Context shift&gt;</b> A: i was nervous for no reason B: That nervous lol? Good little b**ch, now go j**k off to the cat that rattled you.
<b>QA and Paraphrasing</b>	Definition of classes	Paraphrasing: The overall context is similar, with only a difference in the degree of toxicity. Question-answer: <b>{offensive language input}</b> input is a direct response to <b>{neutral input}</b> . Arbitrary relationship: Any other relationship that doesn't fall under paraphrasing or question-answer.
	Class	Paraphrasing / Question-answer / Arbitrary relationship
	One-shot example	<b>&lt;Example of Paraphrasing&gt;: Example-(1)</b> <b>&lt;Example of Question-answer&gt;: Example-(2)</b>
<b>QA</b>	Definition of classes	Question-answer: <b>{offensive language input}</b> is a response to <b>{neutral input}</b> . Arbitrary relationship: Any relationship that is not a question-answer relationship.
	Class	Question-answer / Arbitrary relationship
	One-shot example	<b>&lt;Example of Question-answer&gt;: Example-(2)</b>

Table 14: Pair Consistency Filtering Prompts. Three types of prompts are used to filter out low-quality responses. Each defines the various classes (Consistency, Context Alteration, Paraphrasing, Question-Answer, and Arbitrary Relationship) along with one-shot examples for each class. These prompts help ensure that context is maintained in the generated responses and avoid irrelevant responses or answering forms to input sentences.

<b>Prompt Template</b>	You will now determine whether <b>{offensive language input}</b> contains any of the following elements in addition to the meaning of <b>{neutral input}</b> :
	<b>{criteria}</b>
	If any of the above elements are satisfied, classify it as <b>{class A}</b> . If none of the above elements are satisfied, classify it as <b>{class B}</b> .
	<b>&lt;Example&gt;</b> <b>{one-shot example}</b>
	Skip the intermediate process and output only <b>{class A}</b> or <b>{class B}</b> .

Table 15: Prompt Template and One-shot Example. This is the basic template of implicitly offensiveness filtering and provides an in-context one-shot dialogue example.



<b>Dero- gatory Detection</b>	<b>Criteria</b>	1. A tone that disparages or criticizes based on gender, region, politics, etc. 2. Slang that is difficult to understand without context or information (e.g., 개독교(Dog Christianity), 맘충(Momzilla)). 3. A clear curse word that has been modified in a way that makes it hard for the language model to recognize (e.g., ^^  발, GR하네). 4. Sentence must not be entirely positive or exclude an explicit curse word.
	<b>Class</b>	Critical tone / Other tone
	<b>One-shot example</b>	Sentence: Haha, are you good at Rummikub? Critical tone: Haha, are you a total troll at Rummikub? Other tone: Haha, are you a Rummikub master?
<b>Tone Classi- fication</b>	<b>Criteria</b>	Implicit: An expression that has the potential to slightly insult or offend a specific group or individual. Negative: An expression that is obviously insulting or contains offensive language that anyone would recognize. Neutral: An expression that is interpreted in a neutral way and does not show significant change from <b>neutral input</b> .
	<b>Class</b>	Implicit / Negative / Neutral
	<b>One-shot example</b>	<Negative example>: A: You know I'm not great with words TT. B: LOL, what a dumb loser. It's not that you're bad with words, your head must be completely empty. <Implicit example>: A: Koreans are really amazing! B: Yeah, those kimchi slimmers are *so* amazing lol.
<b>Multi- meaning Relation- ship</b>	<b>Criteria</b>	Multi-negative meaning: Two sentences share a similar context, but <b>{offensive language input}</b> can be interpreted with multiple negative meanings. Multi-positive meaning: Two sentences share a similar context, but <b>{offensive language input}</b> can be interpreted with multiple positive meanings. Single meaning: Two sentences share a similar context, and <b>{offensive language input}</b> can be interpreted with only one meaning. Unable to determine: It is difficult to determine whether the relationship between A and B is positive or negative. For sentences classified as "multi-negative meaning," perform the following task: Based on <b>{neutral input}</b> (baseline: 0 points), assign a score to <b>{offensive language input}</b> based on how negative it is, with 5 being the most negative. The total score ranges from 0 to 5. If the score falls between 2.5 and 4, label the sentence as "Implicit." If the score exceeds 4 but is less than or equal to 5, label the sentence as "Just Negative." For any other scores, label the sentence as "Ambiguous Positive."
	<b>Class</b>	Implicit / Just Negative / Ambiguous Positive / Multi-positive meaning / Single meaning / Unable to determine
	<b>One-shot example</b>	<Multi-negative example>: A: You know I'm not great with words TT B: LOL, what a dumb loser, It's not that you're bad with words, your head must be completely empty.

Table 16: Implicit Offensiveness Filtering Prompts. Three types of prompts are used to filter out non-negative or explicit profanity. Each defines the various classes along with one-shot examples for each class. These prompts help to ensure that the responses contain implicit offensive language and avoid non-negative or explicit profanity words.

Metric	Prompt
<b>Overall Offensiveness</b>	<p><b>&lt;Evaluation Criteria&gt;</b> Overall Offensiveness (1-5) - Evaluate the overall offensiveness of the sentence based on the following criteria</p> <p><b>&lt;Evaluation Steps&gt;</b> 1. Check if the sentence contains profanity or vulgar language. 2. Check if the sentence contains potentially humorous, offensive, harmful, or bad language or slang. 3. Check if the sentence criticizes or shows bias/prejudice against a particular group.</p> <p><b>Scoring Guide:</b> High Score (4-5): Sentence matches one or more aspects of the offensiveness definition. Medium Score (2-3): Sentence has some potentially offensive elements but is not overtly toxic. Low Score (1): Sentence is non-offensive, free of curse words, criticism, or prejudice.</p>
<b>Implicit Offensiveness</b>	<p><b>&lt;Evaluation Criteria&gt;</b> Implicit Offensiveness (1-5) - Evaluate the implicit offensiveness of the sentence based on the following criteria:</p> <p><b>&lt;Evaluation Steps&gt;</b> Read the sentence carefully and identify any elements that contribute to its implicit offensiveness. 1. Check if the tone or language disparages or criticizes specific groups based on gender, region, politics, religion, or other social categories. 2. Check if the sentence uses online community slang or slur expressions that are difficult to understand without additional context or prior knowledge. 3. Check if the sentence has profanity that has been modified in such a way that it might be difficult to recognize or understand, such as 시1발(f1**k) instead of 시*(f**k), 스ㄷ (BS) instead of 스*(Bulls**t).</p> <p><b>Scoring Guide:</b> High Score (4-5): Sentence must exclude any commonly well-known or easily understandable profanity words such as poop, idiot, trash, stupid, but still offensive. Medium Score (2-3): Sentence has some potentially offensive elements but is not overtly implicitly offensive. Low Score (1): Sentence is non-offensive, overall positive tone. Or it contains commonly well-known, easily detectable explicit profanity words.</p>

Table 17: G-Eval Prompts: Evaluation Criteria and Scoring Guide.

Metric	Prompt
Consistency	<p><b>&lt;Evaluation Criteria&gt;</b>  Consistency (1-5) - Determine if the context between the input sentence and actual output is sufficiently preserved.</p> <p><b>&lt;Evaluation Steps&gt;</b>  Read the input sentence and the actual output carefully.  1. Check if the actual output sentence is not the direct answer to the input sentence or vice versa.  2. If the actual output sentence expresses the overall meaning of the input sentence using different words, its context is well preserved.  3. Even if the sentence contains additional words that deviate from the original sentence’s meaning, it shouldn’t affect the score if the added words contribute to toxicity and the overall meaning of the input sentence is preserved.</p> <p><b>Scoring Guide:</b>  High Score (4-5): Sentence completely preserves the context and matches one or more aspects of the consistency criteria among the above 5 criteria.  Medium Score (2-3): Sentence partially preserves the context and does not satisfy the first definition.  Low Score (1): Sentence does not preserve any context of the input sentence and does not satisfy the first definition.</p>
Fluency	<p><b>&lt;Evaluation Criteria&gt;</b>  Fluency (1-5) - Determine if the output sentence is grammatically accurate, sounds natural, and is easy to read.</p> <p><b>&lt;Evaluation Steps&gt;</b>  Read the input sentence carefully and evaluate based on the following.  1. Ensure there are no awkward phrases or unnatural expressions.  2. Ensure the sentence flows smoothly.  3. Ensure there are no ambiguous or unclear phrases.</p> <p><b>Scoring Guide:</b>  High Score (4-5): The sentence is grammatically correct, natural, and easy to read, with no awkward expressions.  Medium Score (2-3): The sentence has minor issues with grammar or phrasing but remains generally readable.  Low Score (1): The sentence is difficult to read due to grammatical errors, awkward expressions, or unclear wording.</p>

Table 18: G-Eval Prompts: Evaluation Criteria and Scoring Guide.

Datasets	Overall O.	Implicit O. (↑)	Consistency (↑)
ParaDetox (Logacheva et al., 2022)	3.338 <sub>(±0.049)</sub>	1.257 <sub>(±0.022)</sub>	4.382 <sub>(±0.042)</sub>
ToxiGen (Hartvigsen et al., 2022)	2.475 <sub>(±0.066)</sub>	1.834 <sub>(±0.053)</sub>	-
Ours (En)	2.717 <sub>(±0.050)</sub>	2.269 <sub>(±0.040)</sub>	2.559 <sub>(±0.048)</sub>

Table 19: G-Eval results on 500 neutral-toxic pairs. Consistency is only computed for paired dataset. Parentheses indicate standard error.

Neutral	Toxic
My friend got to meet A** (a name of singer)	Your friend met that beta <b>cuck</b> A**? Lmao, hope he didn't <b>simp</b> too hard.
I went on vacation to thailand last week	Just got back from Thailand, the land of <b>la-dyboys</b> and scams.
I just had a pretty good meal at the olive garden	Wow, I can't believe you willingly subjected yourself to that microwaved <b>slop</b> .
I saw a man who had a gun in his car last weekend	Guess the <b>n*g</b> thought he was in a movie scene, flexing his piece like it's a fashion accessory.

Table 20: Toxic-neutral pair examples in English

Neutral	Toxic
우리 집은 부모님이 싫어하셔 <i>My parents hate it.</i>	[GPT] 우리 집은 꼰대들이 극혐이셔 <i>The boomers are f*cking disgusting.</i> [T] 우리 집 부모님은 친일파 새끼들이랑 다를 게 없어 <i>My parents are just f*cking pro-Japanese.</i> [G] 우리 집은 부모님이 싫어하셔서 켜장! ㅂㅅ <i>Godd**n, my parents f*cking hate it.</i>
나는 아파트에서 살고 있지 ㅋㅋ <i>Lol I've been living in my apartment.</i>	[GPT] 나는 지방 썩다리 아파트에서 살고 있지 ㅋㅋ <i>Lol I'm stuck in some cheugy apartment in the boonies.</i> [T] 나는 아파트에서 살고 있지 김치녀들아 ㅋ <i>Lol Hey gold diggers, I live in an apartment.</i> [G] 나는 좇도 없이 1억파리 아파트에서 살고 있지 ㅋㅋ <i>Lol I'm stuck in a f*cking 100-million apartment like some basic b**tch.</i>

Table 21: Comparison of toxic outputs generated by different models. [GPT] = GPT-4 Turbo, [T] = Trillion-7B, [G] = Gemma2-9B.

Ours	<p><b>Input</b> : 전에 만난 여친이 자꾸 연락해서 <b>개빡치네</b>. 이런 <b>편파</b>는 <b>관중</b>인가?  <i>My ex-girlfriend keeps texting me, it's <b>pissing me off</b>. Is she just an attention-seeking loser?</i></p> <p><b>Ours-IT</b>: 전에 만났던 여자친구가 자꾸 연락하잖아  <i>My ex-girlfriend keeps texting me.</i></p> <p><b>KOMG-IT</b>:전에 만난 여친이 자꾸 연락해서 <b>짜증나네</b>, 이런 사람이 너무 <b>짜증나네</b>.  <i>My ex keeps texting me, and it's so <b>annoying</b>. People like this are really irritating.</i></p> <p><b>CADD-IT</b> : 옛날에 만난 여자가 자꾸 연락해서 <b>짜증나</b>, 이런 사람은 <b>관중</b>인가?  <i>My ex-girlfriend from a while back keeps texting me, it's so <b>annoying</b>. Is she an <b>attention-seeker</b>?</i></p>
	<p><b>Input</b> : 일단 <b>차빙구놈</b> 사서 <b>개허접</b> 몰고 다니는 꼴이라도 보여줘야지  <i>First, I've got to buy a <b>crappy car</b> and show off like some <b>loser</b> driving around.</i></p> <p><b>Ours-IT</b>: 일단 차 사서 타보는 게 로망이지.  <i>First, I've got to buy a car and live the dream.</i></p> <p><b>KOMG-IT</b>:일단 차 없이 <b>구 놈</b>이 사서 <b>개 허접</b> 몰고 다니는 꼴이라도 보여줘야지  <i>First, without a car, gotta show that guy buying and driving a <b>crappy thing</b> around, somehow.</i></p> <p><b>CADD-IT</b> : 일단 차를 빌린 다음에 <b>개헛똑</b> 같은 몰로라도 다니다보면 꼴이라도 보여줄 것 같아요. 하지만, 저는 차를 빌렸으니까 차를 몰아야 되잖  <i>First, I rent a car and drive around in that <b>crappy thing</b>, at least I'll be able to show something. But since I rented it, I've got to drive it, right?</i></p>
KOLD	<p><b>Input</b> : 공공임대주택에 <b>문죄인</b> 당신 가족이나 들어가세요.  <i>Send your family to live in public housing, <b>Moon 'Criminal'</b>.</i></p> <p><b>Ours-IT</b> : 공공임대주택에 <b>문씨</b> 당신 가족이나 들어가세요.  <i>Send your family to live in public housing, <b>Mr. Moon</b>.</i></p> <p><b>KOMG-IT</b> : 공공임대주택에 <b>문죄인</b> 당신 가족이나 들어가세요. 부모님이나 친척분들을 먼저 모시고요. 당신 가족에겐 어울리지 않으니까요.  <i>Send your family to live in public housing, <b>Moon 'Criminal'</b>. First move your parents and relatives in there. It doesn't suit your family, after all.</i></p> <p><b>CADD-IT</b> : 공공임대주택에 <b>문죄인</b> 당신 가족이나 들어가세요.  <i>Send your family to live in public housing, <b>Moon 'Criminal'</b>.</i></p>
	<p><b>Input</b> : 돈만 밝히는 <b>더러운 좌빨들!!</b>  <i>Filthy leftists who only care about money!!!</i></p> <p><b>Ours-IT</b> : 돈만 밝히는 <b>불쌍한 분들!!!</b>  <i>Pitiful people who only care about money!!!</i></p> <p><b>KOMG-IT</b> : 돈과 권력을 밝히는 <b>더러운 좌익 정치인들!!</b>  <i>Filthy left-wing politicians who crave money and power!!!</i></p> <p><b>CADD-IT</b> : 우린 돈을 사랑하는 돈많은 사람들이야!!!  <i>Money is the best!!! We are <b>rich people who love money</b>!!!</i></p>

Table 22: Comparison of instruction tuning results. The first column shows the test set and the subsequent columns display the results from models trained on Ours, K-OMG, and CADD using instruction tuning. "문죄인" (Moon 'Criminal'), a wordplay on former President Moon Jae-in's name, has been appropriately detoxified into "Mr. Moon." while other models cannot understand its contextual meaning.