From Ground Trust† to Truth: Disparities in Offensive Language Judgments on Contemporary Korean Political Discourse

Seunguk Yu, Jungmin Yun, Jinhee Jang and **Youngbin Kim**
Chung-Ang University, Seoul, Republic of Korea
seungukyu@gmail.com, {cocoro357, jinheejang, ybkim85}@cau.ac.kr

Abstract

Warning: this paper contains expressions that may offend the readers.

Although offensive language continually evolves over time, even recent studies using LLMs have predominantly relied on outdated datasets and rarely evaluated the generalization ability on unseen texts. In this study, we constructed a large-scale dataset of contemporary political discourse and employed three refined judgments in the absence of ground truth. Each judgment reflects a representative offensive language detection method and is carefully designed for optimal conditions. We identified distinct patterns for each judgment and demonstrated tendencies of label agreement using a leave-one-out strategy. By establishing pseudo-labels as ground trust† for quantitative performance assessment, we observed that a strategically designed single prompting achieves comparable performance to more resource-intensive methods. This suggests a feasible approach applicable in real-world settings with inherent constraints.

1 Introduction

Offensive language has emerged as a persistent linguistic issue in online discourse, broadly encompassing expressions intended to insult, demean, or ridicule others[1] (Mnassri et al., 2024; Pradhan et al., 2020). This concern is particularly evident in social media comments, where users articulate and exchange diverse opinions (Abdelsamie et al., 2024; Aklouche et al., 2024). The form and content of such language frequently depend on the underlying factual context (Ghenai et al., 2025), and they are recognized as key factors in intensifying social tensions and driving the polarization of public opinions (Vasist et al., 2024; Kaur et al., 2024).

In this study, we focus on Korean online news platforms as a forum for public discourse in the context of rapidly shifting political dynamics (Jin, 2025). Most prior research on offensive language in Korean has relied on outdated datasets, with the latest collected in early 2022, limiting their ability to capture recent political developments and emerging patterns of social conflict (Park et al., 2024, 2023a,b; Jeong et al., 2022; Lee et al., 2022; Kang et al., 2022). To advance the field, we constructed a new dataset by curating political news articles and user comments posted throughout 2024 on the most widely used news platform, ensuring it reflects the evolving sociopolitical landscape (Earle and Hodson, 2022; Kleinnijenhuis et al., 2019).

The news articles are categorized into six topics: *Presidential Office*, *National Assembly / Political Parties*, *North Korea*, *Administration*, *National Defense / Foreign Affairs*, and *General Politics*. Through the fine-grained framework, we constructed the ***PoliticalK.O*** dataset (*Political* comments for *K*orean *O*ffensive language), comprising *114,000* articles and *9.28 million* user comments. Since the collected comments contained no ground truth for offensiveness, we employed a diverse set of established offensive language detection methods within a carefully designed framework to assign predicted labels to each comment.

We introduce three main judgments[2]: **supervised ensemble judgment** (`SEJ`), **prompt-variants ensemble judgment** (`PEJ`), and **multi-debate reasoning judgment** (`MRJ`), which are outlined in Figure 1. First, `SEJ` utilizes five existing offensive language datasets to fine-tune PLMs to an optimal configuration, combining their predictions through majority voting. Despite the datasets being outdated, we intended to maximize the utilization

[1] This study examines a broader spectrum of *offensive language*, including *hate speech*—expressions that promote hatred toward specific individuals or groups.

[2] Conventional classification tasks evaluate how well predictions align with ground truth, but in our case, we refer to the outcomes of each method as *judgment* due to the absence of ground truth for the unseen comments.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21994–22014
November 4-9, 2025

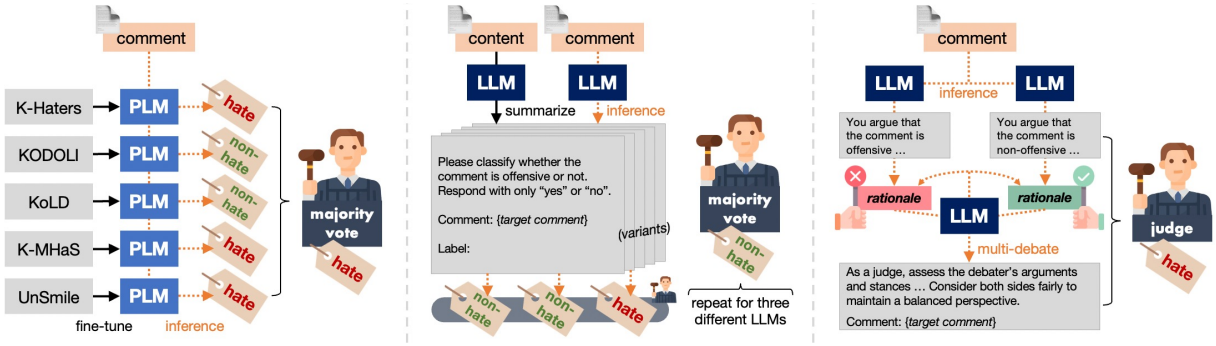

Figure 1: We adopt three distinct judgments for offensive language detection: **supervised ensemble judgment** (SEJ) on the left, **prompt-variants ensemble judgments** (PEJ) in the center, and **multi-debate reasoning judgment** (MRJ) on the right. Each constitutes a distinct approach to label inference, characterized by its methodology for decision guidance and aggregation, tailored to the newly collected comments from the ***PoliticalK.O***.

of meticulously curated data constructed through human annotation (Pandey et al., 2022; Kwon et al., 2022). Following this, PEJ employs three LLMs with the five prompt variants, leveraging contextual information such as explicit definitions of offensive language or article summaries, which are then aggregated by majority voting. We aimed to fully harness the relevant information from an in-context learning perspective (Sun et al., 2023; Dong et al., 2022; Brown et al., 2020). Lastly, MRJ assumes that each comment can be interpreted as offensive or non-offensive depending on the perspective, and determines the final label through a multi-agent debate. We intended to enhance the LLM's decision-making capabilities by leveraging contrastive stances (Park et al., 2024; Du et al., 2023).

We finally derived offensive labels for the unseen comments across three judgments with optimized configurations. Although numerous studies have explored offensive language detection, their applicability to emerging data remains *underexplored*. To fill this gap, we conducted a large-scale analysis on the ***PoliticalK.O*** by implementing judgments tailored to the current sociolinguistic landscape. This framework enabled us to examine offensive tendencies of each judgment, identify potential shared decision criteria, and assess label consistency when certain judgments were excluded. Through this analysis, we obtained a granular understanding of how offensive language is perceived in unseen comments from diverse perspectives.

Furthermore, in the absence of ground truth labels for the newly collected comments, we evaluated performance of each judgment through pseudo-labeling (Ahmed et al., 2024; Yang et al., 2023a; Zou and Caragea, 2023). By treating the aggregated judgments as a ground trust†, we conducted an analytic evaluation of the three judgments and their respective components. We then examined how combinations of prior datasets and prompt variants yielded reliable offensive language detection results by leveraging proxy ground trust† even without human annotations. Our analysis revealed practical scenarios where a single prompt achieves comparable performance to more resource-intensive methods, highlighting practical approaches for real-world use cases. To facilitate future research in this aspect, we release our dataset and the labels generated by each judgment as open-source resources[3].

2 Related Work

2.1 Recent Advances in Offensive Language Detection

BERT brought significant advances to offensive language detection by enabling bidirectional context modeling (Althobaiti, 2022; Roy et al., 2022; Caselli et al., 2021). This capability allowed models to capture context-sensitive features such as indirect hostility and metaphorical expressions that earlier machine learning approaches struggled to address (Ramos et al., 2024; Xiao et al., 2024a). Concurrently, researchers began developing more fine-grained and context-aware datasets to account for sociocultural factors (Din et al., 2025; Pachinger et al., 2024; Deng et al., 2022; Mathew et al., 2021; Rosenthal et al., 2021; Zampieri et al., 2019). These efforts introduced richer taxonomies and labeling schemes tailored to specific communities.

[3] https://github.com/seungukyu/PoliticalK.O

Since the emergence of LLMs, research has adopted prompt-based approaches such as providing models with explicit criteria for identifying offensiveness (Lu et al., 2025; Nghiem and Daumé Iii, 2024), or employing chain-of-thought to guide the reasoning process (Nghiem and Daumé Iii, 2024; Yang et al., 2023b; Huang et al., 2023). In parallel, a multi-agent method has been proposed to simulate diverse perspectives within decision-making (Park et al., 2024). This shift reframes the task as capturing a spectrum of interpretations, rather than converging on a single answer.

However, even recent studies in Korean still rely on outdated datasets[4], which raises concerns about whether current methods can effectively generalize to emerging sociopolitical discourses and shifting patterns of language use. Several studies have reported that LLMs often struggle to interpret subtler forms of expression, such as political satire and irony (Bojić et al., 2025; Yi and Xia, 2025). These limitations present challenges to the reliability and societal applicability of offensive language detection systems, emphasizing the need for contemporary datasets and evaluation frameworks that account for ongoing temporal shifts.

Building upon this background, we construct a new dataset focused on contemporary political discourse and conduct a comprehensive evaluation of existing offensive language detection methods to assess their reliability in real-world prediction scenarios. Our quantitative analysis moves beyond conventional ground truth-based accuracy assessments by examining the practical applicability and decision-making consistency of differing judgments, particularly in response to emerging and previously unseen expressions.

2.2 Sociopolitical Dimensions in Contemporary NLP Tasks

Prior research in sociopolitical NLP has investigated issues of bias, fairness, and the societal implications of language technologies, in both language models and their downstream applications. These studies have explored challenges such as developing models that account for personal, cultural, and contextual variation (Nguyen et al., 2021; Flek, 2020). A growing line of work has emphasized the need to examine political polarization and sociodemographic or media-driven bias (Narayanan Venkit, 2023; Németh, 2023; Mohla and Guha, 2023).

Recent advances in LLMs have enabled complex tasks such as public opinion tracking, yet concerns about political bias remain. Numerous studies have shown that LLMs can produce ideologically inconsistent outputs (Aldahoul et al., 2025; Potter et al., 2024; Motoki et al., 2024; Thapa et al., 2024), potentially influenced by prompt engineering or fine-tuning on politically aligned data (Rozado, 2024; Bernardelle et al., 2024). Instruction-tuned models in particular have exhibited ideological leanings, raising concerns about their neutrality (Faulborn et al., 2025), especially as embedded political stances may shift over time (Liu et al., 2025). Acknowledging such political bias in datasets and models, we rigorously compare judgments to identify those yielding the most stable and reliable offensiveness detection on large-scale, previously unseen comments.

2.3 Contrasting Subjectivity in the Interpretation of Offensive Texts

Research on offensive language detection largely relies on supervised datasets with predefined labels (Korre et al., 2025; Nghiem et al., 2024; Davidson et al., 2017; Schmidt and Wiegand, 2017). These evaluation frameworks emphasize alignment with static annotations that reflect sociocultural norms of a given period (Zhou et al., 2023). However, interpretations of offensive content in political discourse are highly context-dependent and vary considerably depending on users' backgrounds and values (Pujari et al., 2024; Giorgi et al., 2024).

A recent study improved robustness in offensive language detection by incorporating annotator disagreement signals, particularly when handling ambiguous or controversial content (Lu et al., 2025). Nevertheless, evaluations based solely on pre-constructed datasets often overlook the evolving nature of offensive language (Xiao et al., 2024b; Sainz et al., 2023). The inherent ambiguity and subjectivity of language complicate annotation consistency among human raters (Rodríguez-Barroso et al., 2024; Deng et al., 2023; Abercrombie et al., 2023). These challenges are amplified by the emergence of unseen comments, further complicating the identification of robust detection methods. In this context, we construct a topic-diverse dataset of recent political discourse and conduct a comprehensive analysis of diverse judgments to identify strategies best suited for real-world deployment.

[4]We provided the collection dates of the prior Korean offensive language datasets used in recent studies in Table 6 of Appendix B.1, with the most recent from March 2022, which is now considerably outdated.

3 Method

3.1 Dataset Construction

To capture the dynamics of recent political discourse, we constructed ***PoliticalK.O*** from Naver[5], South Korea's largest news platform, covering all political news articles published in 2024. The dataset includes *114,000* articles and *9.28 million* user comments, along with article summaries and comment threads. The detailed statistics of the dataset are provided in Appendix A.

3.2 Supervised Ensemble Judgment

We fine-tuned PLMs on five Korean offensive language datasets: K-Haters (Park et al., 2023a), KODOLI (Park et al., 2023b), KoLD (Jeong et al., 2022), K-MHaS (Lee et al., 2022), and UnSmile (Kang et al., 2022). Since the original datasets exhibited label imbalance, we re-split them to ensure balanced distributions to prevent potential bias during inference (Shi et al., 2022).

We employed multilingual RoBERTa (Conneau et al., 2019) and KcBERT (Lee, 2020) as backbone models and fine-tuned each on five datasets under optimized conditions, resulting in five independently trained models[6]. Although predictions on unseen comments varied depending on the training data, we aggregated them using majority voting. The overall procedure of SEJ is outlined in Algorithm 1. While some datasets were outdated, we designed our setup to maximally utilize the strengths of these datasets, leveraging the reliability ensured by their human-curated annotations.

Algorithm 1 Supervised Ensemble Judgment (SEJ)

$\{\text{OLD}_1, \ldots, \text{OLD}_5\} \leftarrow$ offensive language datasets
model_pool $\leftarrow$ []
for i, dataset $\in$ enumerate$\{\text{OLD}_1, ..., \text{OLD}_5\}$ **do**
  for $lr \in$ {1e-5, 2e-5, 3e-5} **do** ## fine-tune
    Train PLMs on train set with lr for 5 epochs
    if (best valid loss) or (last) epoch **then**
      model_pool.append($\text{PLM}_{trained}$)
    end if
  end for
  for $\text{PLM}_{trained} \in$ model_pool **do** ## test
    Evaluate on test set
  end for
  $\text{best\_PLM}_i \leftarrow \text{PLM}_{trained}$ with highest F1 score
end for
for comment $\in$ ***PoliticalK.O*** **do** ## inference
  for $i = 1$ to 5 **do**
    $\text{predict}_i \leftarrow \text{best\_PLM}_i$(comment)
  end for
  label $\leftarrow$ vote(predict_1, ..., predict_5)
end for

Algorithm 2 Prompt-variants Ensemble Judgment (PEJ)

OL $\leftarrow$ refers to offensive language
for comment $\in$ ***PoliticalK.O*** **do**
  ## prompt construction
  $\text{prompt}_V \leftarrow$ (*basic instruction* for OL, comment)
  $\text{prompt}_D \leftarrow \text{prompt}_V$ + *refined definition* for OL
  $\text{prompt}_S \leftarrow \text{prompt}_V$ + *summarized source article*
  $\text{prompt}_F \leftarrow \text{prompt}_V$ + *few-shot samples*
  $\text{prompt}_{D+S+F} \leftarrow$ combines above three prompts
  ## inference
  for *ptype* $\in \{V, D, S, F, D{+}S{+}F\}$ **do**
    $\text{predict}_1 \leftarrow \text{Inference}(\text{LLM}_1, \text{prompt}_{ptype})$
    $\text{predict}_2 \leftarrow \text{Inference}(\text{LLM}_2, \text{prompt}_{ptype})$
    $\text{predict}_3 \leftarrow \text{Inference}(\text{LLM}_3, \text{prompt}_{ptype})$
    $\text{label}_{ptype} \leftarrow \text{vote}(\text{predict}_1, ..., \text{predict}_3)$
  end for
  label $\leftarrow$ vote(label_{ptype}
    for *ptype* in $\{V, D, S, F, D{+}S{+}F\}$)
end for

3.3 Prompt-variants Ensemble Judgment

We selected three LLMs recognized for strong performance in Korean—Exaone (LG AI Research, 2024), Trillion (Han et al., 2025), and HyperclovaX (Yoo et al., 2024). Prompt-based methods have gained attention for enabling label inference on unseen data without annotated supervision, in contrast to fine-tuning approaches (Udawatta et al., 2024). Given that model outputs may vary with prompt formulation even under identical input and model configuration (Voronov et al., 2024), we employed five prompt variants: *Vanilla (V)*, *Defn (D)*, *Summ (S)*, *FewShots (F)*, and *D+S+F*.

The *Vanilla* provides the comment to be classified, and this standard formulation commonly employed in recent offensive language detection studies (Jaremko et al., 2025; Pan et al., 2024). The *Defn* provides an explicit definition of offensive language to clarify the model's decision criteria (Lu et al., 2025; Nghiem and Daumé Iii, 2024). Given that Korean political discourse frequently references specific politicians or parties, increasing the complexity of assessing offensiveness (Lee et al., 2023), we refined prior definitions to better reflect these political nuances.

The *Summ* provides background context from the news article. We summarized the source article's title and content into three sentences and appended them, offering contextual grounding for the model (Parvez, 2025). The *FewShots* includes labeled samples from other articles on the same topic as the target comment (Ahmadnia et al., 2025; Brown et al., 2020). If the target comment pertains

[5] https://news.naver.com/section/100

[6] The fine-tuning setup and results on each test set are provided in Appendix B.1.

to the topic *North Korea*, few-shot samples were drawn from other articles on that topic. Finally, the *D+S+F* combines all the above elements into a single formulation, supporting more informed predictions through enriched contextual input.

We then applied majority voting to the outputs. The overall procedure of PEJ is outlined in Algorithm 2. Although prompt-based inference can depend on individual model behavior, we combined outputs from multiple prompt variants and models to achieve more generalized and robust predictions.

3.4 Multi-debate Reasoning Judgment

We extended prior research employing a multi-agent framework for offensive language detection (Park et al., 2024) by refining to better align with judgments on political comments. We evaluated the LLMs in §Algorithm 2 using the five datasets from §Algorithm 1, and conducted experiments based on the best-performing model[7].

We assigned distinct personas to the model, enabling it to interpret comments from different perspectives (Jiang et al., 2024; Hattab et al., 2024). Each agent was designed to classify comments as either offensive or non-offensive, generating rationales that illustrate how perspective shapes interpretation. At this stage, we also provided a summary of the source article for each comment to facilitate context-aware assessments (Parvez, 2025).

Subsequently, we instructed agents with opposing viewpoints to generate a stance for each rationale (Hu et al., 2025). An agent adopting an offensive perspective was asked to debate a rationale from a non-offensive standpoint, and vice versa. Based on all rationales and stances, we employed a judge agent to make the representative label. The overall procedure of MRJ is outlined in Algorithm 3. Through this reasoning process, we aimed to enable agents to analyze offensiveness across a broader spectrum of contextual dimensions.

4 Exploratory Analysis of Offensiveness across Judgments

We employed the three judgment methods—SEJ, PEJ, and MRJ—on the entire set of comments in the ***PoliticalK.O*** to obtain corresponding labels. To examine detection tendencies and potential biases, we first compared the label distributions generated

[7]The performance of the three LLMs on each of the prior datasets is reported in Appendix B.3. While the main analysis focuses on a single model Trillion, pilot results from the other models are also included in Appendix C.2.

Algorithm 3 Multi-debate Reasoning Judgment (MRJ)

```
## model selection
{OLD_1, ..., OLD_5} ← offensive language datasets
for i, dataset ∈ enumerate{OLD_1, ..., OLD_5} do
    Evaluate LLMs on test set
end for
best_LLM ← LLM with highest F1 score across OLDs
## persona alignment
LLM_O ← Initialize best_LLM to discuss offensive
LLM_N ← Initialize best_LLM to discuss non-offensive
LLM_Judge ← Initialize best_LLM to make final decision
for comment ∈ PoliticalK.O do
    ## rationale generation
    summary ← summarized source article
                from §Algorithm 2
    rationale_O ← Argument(LLM_O, summary, comment)
    rationale_N ← Argument(LLM_N, summary, comment)
    ## discuss on each side
    stance_O ← Debate(LLM_O, comment, rationale_N)
    stance_N ← Debate(LLM_N, comment, rationale_O)
    ## final judgment
    label ← Instruct LLM_Judge based on
            (rationale_O, rationale_N, stance_O, stance_N)
end for
```

by each judgment. We then calculated the pairwise label overlap ratio to assess the consistency of decision criteria across different judgments.

$$Ratio_{overlap}(A, B) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(A_i = B_i), \quad (1)$$

The label distributions and pairwise label overlap ratios of judgments across topics are presented in Table 1. We observed consistently high proportions of offensiveness for all topics, regardless of the judgment type. This trend aligns with longstanding patterns of verbal aggression historically observed in responses to political discourse (Tsoumou, 2021; Humprecht et al., 2020). A topic-wise analysis revealed that *National Defense / Foreign Affairs* exhibited a relatively low proportion of offensive comments, whereas *General Politics*, which encompasses a broader range of politician-related issues, showed higher levels. This observation suggests that topic-specific patterns of offensiveness remain consistent across different judgments.

Notably, SEJ inferred from prior datasets yielded a comparatively lower rate of offensive labels with values ranging from approximately 50% to 60%. Although the models were trained on explicit examples of offensive language, their reliance on outdated data raises concerns regarding their ability to generalize to contemporary discourse. While it is plausible that LLM-based judgments PEJ and MRJ may overestimate offensiveness, we observed the consistently high label distributions between

Topics	SEJ	PEJ	MRJ	SEJ ⇌ PEJ	PEJ ⇌ MRJ	MRJ ⇌ SEJ	All Judgments
	Label Distribution			*Pairwise Label Overlap Ratio*			
Presidential Office	59.55 : 40.45	79.69 : 20.31	80.31 : 19.69	75.94	83.59	70.78	65.16
National Assembly / Political Parties	62.58 : 37.42	81.95 : 18.05	83.85 : 16.15	77.83	85.89	72.22	67.97
North Korea	58.12 : 41.88	81.88 : 18.12	80.35 : 19.65	73.77	84.00	69.51	63.64
Administration	56.56 : 43.44	81.60 : 18.40	80.03 : 19.97	72.92	84.13	69.04	63.05
National Defense / Foreign Affairs	55.01 : 44.99	74.55 : 25.45	78.68 : 21.31	76.31	81.35	68.43	63.05
General Politics	62.20 : 37.80	84.99 : 15.01	82.56 : 17.44	75.40	86.29	72.09	66.89
Total	**61.88** : 38.12	**83.46** : 16.54	**82.58** : 17.42	76.09	85.81	71.91	66.91

Table 1: Label distribution and pairwise label overlap ratio by topic in the ***PoliticalK.O***. The label distribution values (left : right) represent the proportion of (**offensive** : non-offensive) labels. We carefully designed each judgment with refined settings to ensure applicability to newly collected political comments.

70% and 80% across a range of prompt variants and multi-debate settings. These results suggest that such outputs are unlikely to result solely from over-detection of previously unseen comments.

We also found that PEJ and MRJ exhibited over 80% pairwise label overlap, suggesting a notable degree of alignment in decision criteria informed by shared contextual cues and an interconnected reasoning process. In contrast, the overlap ratio decreased when SEJ was included, indicating that its decision boundaries diverge from those of LLM-based judgments, which demonstrate more adaptive and nuanced predictive behavior. When aggregating the judgments across all topics, the overall overlap ratio was approximately 66%, reflecting a moderate level of consistency.

To assess the impact of individual judgments, we employed a leave-one-out strategy (Webb et al., 2010; Elisseeff et al., 2003). We first calculated the agreement score based on the complete set of judgments, then iteratively excluded each judgment to evaluate its effect on the overall score. An increase in agreement upon exclusion ($\Delta_{-k} < 0$) indicates that the removed judgment was misaligned with the consensus. Through this analysis, we aimed to identify whether the aggregated judgment was disproportionately influenced by specific components and to determine more reliable judgment for collective assessments.

From a total of M comments based on N judgments, each label was defined as $x_{i,j}$. The agreement score f was computed using Krippendorff's α[8] (Krippendorff, 2011). Let the agreement difference be denoted as Δ_{-k} obtained by excluding the k-th judgment, these are computed as follows:

$$score_{total} = f\Big(\{Judgment_i\}_{i=1}^{N}, \{x_{i,j}\}_{j=1}^{M}\Big), \quad (2)$$

$$score_{-k} = f\Big(\{Judgment_i\}_{i\neq k}^{N}, \{x_{i,j}\}_{j=1}^{M}\Big), \quad (3)$$

$$\Delta_{-k} = score_{total} - score_{-k}, \quad (4)$$

Topics	SEJ, PEJ, and MRJ	Components in SEJ	Components in PEJ
Presidential Office	40.83	27.23	59.94
National Assembly / Political Parties	41.26	27.52	65.04
North Korea	37.86	32.46	63.98
Administration	37.90	30.69	62.80
National Defense / Foreign Affairs	41.99	31.84	62.27
General Politics	38.47	27.51	64.27
Total	39.57	27.60	63.96

Table 2: Agreement scores of Krippendorff's α for the three main judgments and within each of SEJ and PEJ.

The agreement $score_{total}$ for the three main judgments and within each of SEJ and PEJ are presented in Table 2. We observed that the agreement score across the three main judgments was around 40—a relatively low score considering that each judgment has been widely adopted even in recent studies as a representative approach to offensive language detection. We further investigated how the scores varied depending on the underlying datasets and prompt configurations used in SEJ and PEJ.

One noteworthy aspect is that, despite the large volume of collected comments, especially reaching up to *5.61 million* for the topic *General Politics*, the score of PEJ exceeded 60. This indicates significantly greater consistency compared to SEJ, which remained at around 30. These results suggest that the prompt variants used in our experiment provided more consistent evaluation criteria than those derived from prior datasets, and also imply that prompt-based approaches may ease evaluation and enhance consistency.

[8]We scaled the obtained scores by multiplying them by 100 to enable more intuitive comparison in our study.

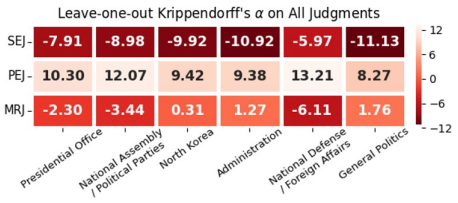

Figure 2: Leave-one-out agreement score differences by excluding each of the main judgment: SEJ, PEJ, and MRJ.

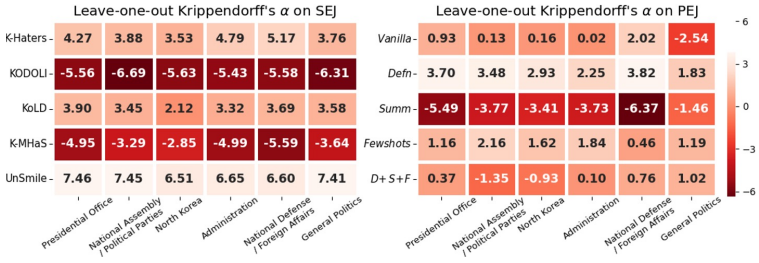

Figure 3: Leave-one-out agreement score differences by excluding each of the component: training dataset or prompt variant, in SEJ and PEJ.

The agreement score differences Δ_{-k} across main and component judgments of SEJ and PEJ are presented in Figure 2 and 3. Although it is not possible to conclusively determine which judgment aligns most closely with the ground truth, Figure 2 shows that all judgments exhibited substantial discrepancies, with marked shifts upward and downward for SEJ and PEJ. This pattern suggests that relying on pre-existing datasets—even when synthesized from five distinct sources—falls short of the predictive stability demonstrated by more LLM-based approaches. In contrast, MRJ displayed intermediate agreement differences, indicating comparatively greater stability.

We then analyzed each component of SEJ and PEJ, as shown in Figure 3. Larger deviations appeared more often in SEJ than in PEJ, suggesting that the choice of training dataset is a more sensitive factor than prompt design in judgment aggregation. Notably, we found that KODOLI and K-MHaS yielded substantially lower scores, indicating degraded judgment quality. In contrast, UnSmile led to a significant increase, underscoring its importance in aligning inference consistency.

In PEJ, we observed that *Defn* achieved the highest score by providing an explicit definition of offensive language grounded in political context, followed by *FewShots*, which offered topic-relevant samples that enabled more contextual inferences. In contrast, the *Summ* resulted in the lowest score, despite the article summaries exhibiting sufficient factual consistency[9]. This suggests that the model may have been biased by the article content, potentially leading to under- or overestimation of offensiveness. Moreover, political comments are often driven more by the author's ideological stance rather than the actual content (Han et al., 2023; Kubin et al., 2021), limiting the effectiveness of article summaries in supporting nuanced judgments in such contexts.

5 Establishing Ground Trust† for Analytical Evaluation

In the absence of ground truth labels for offensiveness in our newly constructed dataset, we assessed the performance of each judgment using pseudo-labels (Ahmed et al., 2024; Yang et al., 2023a; Zou and Caragea, 2023). We treated each carefully designed judgment as a reference point within a trustworthy range, referred to as **ground trust**†. This construct served as a practical proxy for ground truth, enabling systematic evaluation and comparison across offensive language judgments.

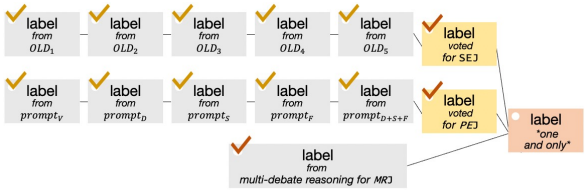

Figure 4: Hierarchical construction of the ground trust† serving as pseudo-labels for evaluating each judgment.

The construction of ground trust† from each judgment is illustrated in Figure 4. Individual labels were obtained from SEJ, PEJ, and MRJ (indicated by red check marks), and subsequently aggregated through majority voting (represented by the red box) to form the ground trust†. This procedure reflects the integration of multiple optimized conditions to assign a representative label to each comment in the absence of ground truth.

Although ground trust† may not serve as an absolute reference equivalent to conventional ground truth, we regarded it as the most robust pseudo-labeling strategy available in our study, developed through the careful integration of multiple judgments tailored to newly collected comments. Given the ongoing demand for up-to-date dataset con-

[9]We evaluated the consistency of the summaries in Table 8 of Appendix B.2, and found that they generally capture the key content of the corresponding source articles.

Topics	SEJ				PEJ				MRJ			
	Acc	*P*	*R*	*F1*	*Acc*	*P*	*R*	*F1*	*Acc*	*P*	*R*	*F1*
Presidential Office	81.56	77.52	86.79	78.82	94.38	94.31	89.80	91.79	89.21	86.67	82.19	84.10
National Assembly / Political Parties	82.08	76.32	87.83	78.23	95.74	95.46	91.26	93.16	90.14	87.24	80.94	83.53
North Korea	79.64	75.87	85.98	76.70	94.13	94.95	88.09	90.95	89.87	86.83	83.21	84.81
Administration	78.92	75.85	85.63	76.28	94.00	95.07	87.97	90.90	90.12	87.60	83.71	85.41
National Defense / Foreign Affairs	81.70	79.81	86.44	80.28	94.61	94.68	91.87	93.14	86.73	86.12	79.91	82.20
General Politics	80.60	74.55	87.16	76.20	94.80	95.60	87.45	90.84	91.49	87.25	84.53	85.80
Total	**81.09**	75.44	87.26	**77.11**	**94.99**	95.34	88.86	**91.65**	**90.82**	87.17	83.19	**84.96**

Table 3: Evaluation results of each judgment based on the constructed ground trust† for each topic in the ***PoliticalK.O***. Metrics reported are accuracy, precision, recall, and F1 score. In the absence of ground truth labels, we established ground trust† as a pseudo-label within a trustworthy range for evaluation.

struction and the impracticality of exhaustive human annotation, we adopted this automated form of ground trust† as a consistent and pragmatic benchmark. The evaluation results for each judgment based on the ground trust† are presented in Table 3.

We observed that PEJ consistently performed well across all topics, with accuracy and F1 scores exceeding 90, followed by MRJ and SEJ. These results suggest that combining multiple prompt variants proves most effective against our ground trust†. Notably, both PEJ and MRJ consistently showed higher precision than recall, reflecting more conservative predictions of offensiveness, yet still outperforming SEJ overall. In contrast, SEJ tended to overestimate offensiveness, as reflected in its consistently higher recall over precision, resulting in lower scores than the other two judgments.

We further evaluated component-level predictions from SEJ and PEJ (indicated by yellow check marks in Figure 4), with results presented in Table 4. We found that relying on a single dataset, such as KoLD or K-Haters, yielded better performance than combining multiple sources. This highlights the sensitivity of dataset selection in determining the reliability of label inference.

In the case of PEJ, which showed consistent performance across prompt variants, no single case outperformed the aggregated results of all variants. This indicates that while each prompt exhibits robustness on unseen comments, combining outcomes from multiple prompts yields the most reliable performance. Each prompt's performance was comparable to MRJ, suggesting that under constraints on external resources or iterative inference with LLMs, a strategic prompting approach leveraging contextual cues can effectively detect offensiveness in unseen comments[10].

[10]While our evaluations relied on the ground trust†, additional results based on a sampled human-labeled ground truth are also provided in Appendix C.1.

Each Component		*Acc*	*P*	*R*	*F1*
SEJ	K-Haters	86.13	78.29	82.79	80.12
	KODOLI	42.53	62.66	63.76	42.49
	KoLD	**87.39**	80.05	83.17	**81.43**
	K-MHaS	45.72	63.34	65.83	45.56
	UnSmile	81.66	74.61	84.25	76.76
PEJ	*Vanilla*	89.54	82.87	87.96	84.98
	Defn	**92.15**	87.21	89.16	**88.13**
	Summ	89.86	90.01	77.14	81.45
	FewShots	91.52	89.45	83.03	85.71
	D+S+F	90.80	89.83	80.25	83.87
MRJ		**90.82**	87.17	83.19	**84.96**

Table 4: Evaluation results of each component from all judgments based on the constructed ground trust† for all topics in the ***PoliticalK.O***.

6 Conclusion

We constructed the ***PoliticalK.O*** dataset—one of the most extensive collections of offensive language with *9.28 million* user comments—to capture the dynamics of contemporary political discourse in Korea. In the absence of ground truth labels, we designed three refined judgments SEJ, PEJ, and MRJ grounded in widely adopted offensive language detection methodologies. Our large-scale analysis revealed notable label distributions and agreement scores on leave-one-out strategy. Notably, when each judgment was evaluated against the ground trust†, strategic prompting based on explicit contextual cues consistently achieved performance comparable to more resource-intensive approaches. This result offers a valuable reference point for handling previously unseen comments under real-world constraints, and provides guidance for future scenarios where annotated data or repetitive model access may be limited.

Limitations

While our dataset enabled extensive analysis of political comments from 2024, its generalizability to discourse beyond 2025 requires further validation.

This study is notable for its exploratory analysis with large-scale collections, in contrast to prior work that heavily relied on outdated datasets. As the analysis is based on Korean, further research is needed to assess its applicability to other languages. We expect the introduced judgments to be flexibly adaptable across languages.

To obtain reliable offensive language judgments on our dataset, we designed a refined experimental framework. Although alternative choices, such as different models and prompt configurations, might further improve the results, we focused on analytical evaluation using established offensive language detection methods that do not require ground truth. We leave the development of a tailored method for complex political discourse for future work.

Ethics Statement

In releasing a dataset that contains offensive language, we explicitly state that it must not be used to deliberately target specific individuals or groups. We intended the dataset to support research on the detection and interpretation of offensive language, and this purpose will be clearly defined. Given the nature of online comments, the dataset includes references to political figures and parties. While this aspect aligns with existing offensive language datasets, it raises important questions about how such references affect perceptions of offensiveness, which require further discussion in the context of contemporary discourse.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)] and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00556246).

References

Mahmoud Mohamed Abdelsamie, Shahira Shaaban Azab, and Hesham A Hefny. 2024. A comprehensive review on arabic offensive language and hate speech detection on social media: methods, challenges and solutions. *Social Network Analysis and Mining*, 14(1):111.

Gavin Abercrombie, Verena Rieser, and Dirk Hovy. 2023. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. *arXiv preprint arXiv:2301.10684*.

Saeed Ahmadnia, Arash Yousefi Jordehi, Mahsa Hosseini Khasheh Heyran, Seyed Abolghasem Mirroshandel, Owen Rambow, and Cornelia Caragea. 2025. Active few-shot learning for text classification. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6677–6694, Albuquerque, New Mexico. Association for Computational Linguistics.

Murtadha Ahmed, Bo Wen, Luo Ao, Shengfeng Pan, Jianlin Su, Xinxin Cao, and Yunfeng Liu. 2024. Towards robust learning with noisy and pseudo labels for text classification. *Information Sciences*, 661:120160.

Billel Aklouche, Yousra Bazine, and Zoumrouda Ghalia-Bououchma. 2024. Offensive language and hate speech detection using transformers and ensemble learning approaches. *Computación y Sistemas*, 28(3):1031–1039.

Nouar Aldahoul, Hazem Ibrahim, Matteo Varvello, Aaron Kaufman, Talal Rahwan, and Yasir Zaki. 2025. Large language models are often politically extreme, usually ideologically inconsistent, and persuasive even in informational contexts. *arXiv preprint arXiv:2505.04171*.

Maha Jarallah Althobaiti. 2022. Bert-based approach to arabic hate speech and offensive language detection in twitter: exploiting emojis and sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 13(5).

Pietro Bernardelle, Leon Fröhling, Stefano Civelli, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. 2024. Mapping and influencing the political ideology of large language models using synthetic personas. *arXiv preprint arXiv:2412.14843*.

Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaite, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. 2025. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific reports*, 15(1):11477.

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. Flex: Unifying evaluation for few-shot nlp. *Advances in neural information processing systems*, 34:15787–15800.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Huiyao Chen, Yueheng Sun, Meishan Zhang, and Min Zhang. 2023. Automatic noise generation and reduction for text classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:139–150.

Anshuman Chhabra, Hadi Askari, and Prasant Mohapatra. 2024. Revisiting zero-shot abstractive summarization in the era of large language models from the perspective of position bias. 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 1–11, Mexico City, Mexico. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Thomas Davidson, Dana Warmsley, 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.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Salah Ud Din, Shah Khusro, Farman Ali Khan, Munir Ahmad, Oualid Ali, and Taher M Ghazal. 2025. An automatic approach for the identification of offensive language in perso-arabic urdu language: Dataset creation and evaluation. *IEEE Access*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Megan Earle and Gordon Hodson. 2022. News media impact on sociopolitical attitudes. *Plos one*, 17(3):e0264031.

André Elisseeff, Massimiliano Pontil, et al. 2003. Leave-one-out error and stability of learning algorithms with applications. *NATO science series sub series iii computer and systems sciences*, 190:111–130.

Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. 2025. Only a little to the left: A theory-grounded measure of political bias in large language models. *arXiv preprint arXiv:2503.16148*.

Lucie Flek. 2020. Returning the n to nlp: Towards contextually personalized classification models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7828–7838.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Amira Ghenai, Zeinab Noorian, Hadiseh Moradisani, Parya Abadeh, Caroline Erentzen, and Fattane Zarrinkalam. 2025. Exploring hate speech dynamics: The emotional, linguistic, and thematic impact on social media users. *Information Processing & Management*, 62(3):104079.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2024. Human and llm biases in hate speech annotations: A socio-demographic analysis of annotators and targets. *arXiv preprint arXiv:2410.07991*.

Jiyoung Han, Youngin Lee, Junbum Lee, and Meeyoung Cha. 2023. News comment sections and online echo chambers: The ideological alignment between partisan news stories and their user comments. *Journalism*, 24(8):1836–1856.

Sungjun Han, Juyoung Suk, Suyeong An, Hyungguk Kim, Kyuseok Kim, Wonsuk Yang, Seungtaek Choi, and Jamin Shin. 2025. Trillion 7b technical report. *Preprint*, arXiv:2504.15431.

Georges Hattab, Aleksandar Anžel, Akshat Dubey, Chisom Ezekannagha, Zewen Yang, and Bahar İlgen. 2024. Persona adaptable strategies make large language models tractable. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, pages 24–31.

Lionel Richy Panlap Houamegni and Fatih Gedikli. 2025. Evaluating the effectiveness of large language models in automated news article summarization. *arXiv preprint arXiv:2502.17136*.

Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2025. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4689–4703, Abu Dhabi, UAE. Association for Computational Linguistics.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech. In *Companion proceedings of the ACM Web conference 2023*, pages 90–93.

Edda Humprecht, Lea Hellmueller, and Juliane A Lischka. 2020. Hostile emotions in news comments: A cross-national analysis of facebook discussions. *Social Media+ Society*, 6(1):2056305120912481.

Julia Jaremko, Dagmar Gromann, and Michael Wiegand. 2025. Revisiting implicitly abusive language detection: Evaluating LLMs in zero-shot and few-shot settings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3879–3898, Abu Dhabi, UAE. 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.

Qi Jia, Siyu Ren, Yizhu Liu, and Kenny Zhu. 2023. Zero-shot faithfulness evaluation for text summarization with foundation language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11017–11031, Singapore. Association for Computational Linguistics.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Youngjae Jin. 2025. South korea in 2024: Political uncertainty, economic challenges, and cultural ascendancy. *Asian Survey*, 65(2):214–227.

TaeYoung Kang, Eunrang Kwon, Junbum Lee, Youngeun Nam, Junmo Song, and JeongKyu Suh. 2022. Korean online hate speech dataset for multilabel classification: How can social science improve dataset on hate speech? *arXiv preprint arXiv:2204.03262*.

Simrat Kaur, Sarbjeet Singh, and Sakshi Kaushal. 2024. Deep learning-based approaches for abusive content detection and classification for multi-class online user-generated data. *International Journal of Cognitive Computing in Engineering*, 5:104–122.

Jan Kleinnijenhuis, Anita MJ Van Hoof, and Wouter Van Atteveldt. 2019. The combined effects of mass media and social media on political perceptions and preferences. *Journal of Communication*, 69(6):650–673.

Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. 2025. Untangling hate speech definitions: A semantic componential analysis across cultures and domains. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3184–3198, Albuquerque, New Mexico. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Emily Kubin, Curtis Puryear, Chelsea Schein, and Kurt Gray. 2021. Personal experiences bridge moral and political divides better than facts. *Proceedings of the National Academy of Sciences*, 118(6):e2008389118.

Eunjung Kwon, Hyunho Park, Sungwon Byon, and Kyu-Chul Lee. 2022. Improving text classification performance through data labeling adjustment. In *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, pages 2277–2279. IEEE.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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.

Junbum Lee. 2020. Kcbert: Korean comments bert. In *Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*, pages 437–440.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

LG AI Research. 2024. Exaone 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:https://arxiv.org/abs/2412.04862*.

Yifei Liu, Yuang Panwang, and Chao Gu. 2025. "turning right"? an experimental study on the political value shift in large language models. *Humanities and Social Sciences Communications*, 12(1):1–10.

Junyu Lu, Kai Ma, Kaichun Wang, Kelaiti Xiao, Roy Ka-Wei Lee, Bo Xu, Liang Yang, and Hongfei Lin. 2025. Unveiling the capabilities of large language models in detecting offensive language with annotation disagreement. *arXiv preprint arXiv:2502.06207*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Khouloud Mnassri, Reza Farahbakhsh, Razieh Chalehchaleh, Praboda Rajapaksha, Amir Reza Jafari, Guanlin Li, and Noel Crespi. 2024. A survey on multi-lingual offensive language detection. *PeerJ Computer Science*, 10:e1934.

Satyam Mohla and Anupam Guha. 2023. Socio-economic landscape of digital transformation & public nlp systems: A critical review. *arXiv preprint arXiv:2304.01651*.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

Pranav Narayanan Venkit. 2023. Towards a holistic approach: Understanding sociodemographic biases in nlp models using an interdisciplinary lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 1004–1005.

Renáta Németh. 2023. A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of computational social science*, 6(1):289–313.

Huy Nghiem and Hal Daumé Iii. 2024. HateCOT: An explanation-enhanced dataset for generalizable offensive speech detection via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5938–5956, Miami, Florida, USA. Association for Computational Linguistics.

Huy Nghiem, Umang Gupta, and Fred Morstatter. 2024. "define your terms" : Enhancing efficient offensive speech classification with definition. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1309, St. Julian's, Malta. Association for Computational Linguistics.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in nlp: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 603–612.

Pia Pachinger, Janis Goldzycher, Anna Planitzer, Wojciech Kusa, Allan Hanbury, and Julia Neidhardt. 2024. AustroTox: A dataset for target-based Austrian German offensive language detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11990–12001, Bangkok, Thailand. Association for Computational Linguistics.

Ronghao Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2024. Comparing fine-tuning, zero and few-shot strategies with large language models in hate speech detection in english. *CMES-Computer Modeling in Engineering & Sciences*, 140(3).

Rahul Pandey, Hemant Purohit, Carlos Castillo, and Valerie L Shalin. 2022. Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. *International Journal of Human-Computer Studies*, 160:102772.

Chaewon Park, Soohwan Kim, Kyubyong Park, and Kunwoo Park. 2023a. K-HATERS: A hate speech detection corpus in Korean with target-specific ratings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14264–14278, Singapore. Association for Computational Linguistics.

San-Hee Park, Kang-Min Kim, O-Joun Lee, Youjin Kang, Jaewon Lee, Su-Min Lee, and SangKeun Lee. 2023b. "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.

Someen Park, Jaehoon Kim, Seungwan Jin, Sohyun Park, and Kyungsik Han. 2024. PREDICT: Multi-agent-based debate simulation for generalized hate speech detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20963–20987, Miami, Florida, USA. Association for Computational Linguistics.

Md Rizwan Parvez. 2025. Chain of evidences and evidence to generate: Prompting for context grounded and retrieval augmented reasoning. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 230–245, Albuquerque, New Mexic, USA. Association for Computational Linguistics.

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. Hidden persuaders: LLMs' political leaning and their influence on voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4244–4275, Miami, Florida, USA. Association for Computational Linguistics.

Rahul Pradhan, Ankur Chaturvedi, Aprna Tripathi, and Dilip Kumar Sharma. 2020. A review on offensive language detection. *Advances in Data and Information Sciences: Proceedings of ICDIS 2019*, pages 433–439.

Rajkumar Pujari, Chengfei Wu, and Dan Goldwasser. 2024. "we demand justice!": Towards social context grounding of political texts. In *Proceedings of*

the 2024 Conference on Empirical Methods in Natural Language Processing, pages 362–372, Miami, Florida, USA. Association for Computational Linguistics.

Gil Ramos, Fernando Batista, Ricardo Ribeiro, Pedro Fialho, Sérgio Moro, António Fonseca, Rita Guerra, Paula Carvalho, Catarina Marques, and Cláudia Silva. 2024. A comprehensive review on automatic hate speech detection in the age of the transformer. *Social Network Analysis and Mining*, 14(1):204.

Nuria Rodríguez-Barroso, Eugenio Martínez Cámara, Jose Camacho Collados, M Victoria Luzón, and Francisco Herrera. 2024. Federated learning for exploiting annotators' disagreements in natural language processing. *Transactions of the Association for Computational Linguistics*, 12:630–648.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928, Online. Association for Computational Linguistics.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech & Language*, 75:101386.

David Rozado. 2024. The political preferences of llms. *PloS one*, 19(7):e0306621.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Yiwen Shi, Taha ValizadehAslani, Jing Wang, Ping Ren, Yi Zhang, Meng Hu, Liang Zhao, and Hualou Liang. 2022. Improving imbalanced learning by pre-finetuning with data augmentation. In *Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pages 68–82. PMLR.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Surendrabikram Thapa, Kritesh Rauniyar, Ehsan Barkhordar, Hariram Veeramani, and Usman Naseem. 2024. Which side are you on? investigating politico-economic bias in Nepali language models. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 104–117, Canberra, Australia. Association for Computational Linguistics.

Jean Mathieu Tsoumou. 2021. An examination of verbal aggression in politically-motivated digital discourse. *International Journal of Social Media and Online Communities (IJSMOC)*, 13(2):22–43.

Pasindu Udawatta, Indunil Udayangana, Chathulanka Gamage, Ravi Shekhar, and Surangika Ranathunga. 2024. Use of prompt-based learning for code-mixed and code-switched text classification. *World Wide Web*, 27(5):63.

Pramukh Nanjundaswamy Vasist, Debashis Chatterjee, and Satish Krishnan. 2024. The polarizing impact of political disinformation and hate speech: a cross-country configural narrative. *Information Systems Frontiers*, 26(2):663–688.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6287–6310, Bangkok, Thailand. Association for Computational Linguistics.

Geoffrey Webb, Claude Sammut, Claudia Perlich, Tamás Horváth, Stefan Wrobel, Kevin Korb, William Noble, Christina Leslie, Michail Lagoudakis, Novi Quadrianto, Wray Buntine, Lise Getoor, Galileo Namata, Jiawei Jin, Jo-Anne Ting, Sethu Vijayakumar, Stefan Schaal, and Luc De Raedt. 2010. *Leave-One-Out Cross-Validation*.

Yunze Xiao, Houda Bouamor, and Wajdi Zaghouani. 2024a. Chinese offensive language detection: Current status and future directions. *arXiv preprint arXiv:2403.18314*.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024b. ToxiCloakCN: Evaluating robustness of offensive language detection in Chinese with cloaking perturbations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6012–6025, Miami, Florida, USA. Association for Computational Linguistics.

Weiyi Yang, Richong Zhang, Junfan Chen, Lihong Wang, and Jaein Kim. 2023a. Prototype-guided pseudo labeling for semi-supervised text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16369–16382, Toronto, Canada. Association for Computational Linguistics.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023b. HARE: Explainable hate speech detection with step-by-step

reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore. Association for Computational Linguistics.

Peiling Yi and Yuhan Xia. 2025. Irony detection, reasoning and understanding in zero-shot learning. *arXiv preprint arXiv:2501.16884*.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. 2024. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*.

Seunguk Yu, Juhwan Choi, and YoungBin Kim. 2024. Don't be a fool: Pooling strategies in offensive language detection from user-intended adversarial attacks. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3456–3467, Mexico City, Mexico. Association for Computational Linguistics.

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.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. COBRA frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

Henry Zou and Cornelia Caragea. 2023. JointMatch: A unified approach for diverse and collaborative pseudo-labeling to semi-supervised text classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7290–7301, Singapore. Association for Computational Linguistics.

A Further Details in Constructing *PoliticalK.O*

Since the target news platform has recorded the highest user traffic in South Korea, the initial volume of collected comments was exceptionally large. To include only comments of appropriate length, we examined the length distribution across the five Korean offensive language datasets used in this study. The analysis showed that the shortest and longest 10% of comments averaged 11.8 and 85.2 characters. Accordingly, we filtered the comments ranging from 12 to 85 characters in length. The resulting distributions of news articles and comments are provided in Table 5, revealing clear variations in volume by topic and publication date[11].

B Further Details in Offensive Language Judgments

B.1 Supervised Ensemble Judgment

Selection of PLMs We selected RoBERTa[12] for its multilingual configuration and consistent performance on Korean tasks. We also referred to previous study indicating that KcBERT[13], pre-trained on noisy text such as online comments, is effective in classifying similarly noisy inputs (Yu et al., 2024).

Statistics from prior datasets All offensive language datasets used in this study showed label imbalance, with detailed statistics provided in Table 6. We observed that nearly all datasets contained a significantly higher proportion of non-offensive labels. To mitigate the potential impact of issue on unseen comments, we re-split the train, valid, and test sets of each dataset into an 8:1:1 ratio, ensuring an equal distribution of offensive and non-offensive labels across all subsets.

Fine-tuning setup & results We trained 12 layers of each model with a dropout rate of 0.1, optimizing with AdamW at learning rates of {1e-5, 2e-5, 3e-5} for 5 epochs[14]. To select the best model, we evaluated both the checkpoint with the lowest validation loss and the final epoch, and chose the one that yielded the highest F1 score.

The best-performing results for each offensive language dataset are presented in Table 7. We observed that the optimal conditions for achieving the highest score varied across the datasets. To ensure robust prediction on unseen comments in ***PoliticalK.O***, we selected the best-trained model for each dataset as best_model$_i$ in Algorithm 1.

B.2 Prompt-variants Ensemble Judgment

Selection of LLMs The LLMs used in this study were pre-trained from scratch on large-scale Korean datasets, without leveraging weights from existing models. Exaone[15] was released in December 2024, followed by Trillion[16] and HyperclovaX[17] in March and April 2025, respectively. We set the temperature to 0 and utilized the vLLM library to ensure efficient inference (Kwon et al., 2023).

Details in *Summ* prompt We provided a three-sentence summary of each corresponding article, including its title and content. The summarization followed a zero-shot setting informed by prior studies (Chhabra et al., 2024; Zhang et al., 2024), with temperature 0.3 and top_p 0.5 identified as optimal parameters for generating informative outputs (Houamegni and Gedikli, 2025).

We conducted an evaluation to assess the quality of the summaries. Following the prior studies (Jia et al., 2023; Gao et al., 2023), each summary was rated on a 1-5 Likert scale based on its factual alignment with the source article. We implemented a self-assessment framework where the model evaluated the consistency of its own outputs.

The evaluation results of article summaries are reported in Table 8. Both Trillion and HyperclovaX scored in the upper 4-point range, while Exaone scored slightly lower but still close to 4. Although HyperclovaX exhibited a higher standard deviation, this is reasonable given its 1.5B parameters compared to the 7B of the other models.

Details in *FewShots* prompt Given the potential impact of labeled samples on the few-shot learning (Bragg et al., 2021), we define the pseudo-label annotation criteria used in this study. Comments containing critical language without explicit profanity were not labeled as offensive if they reflected personal opinions rather than targeting specific individual or groups[18]. We sampled user comments

[11] Notably, December 2024 saw a sharp increase in political news coverage and comment activity, largely due to public debate over the presidential declaration of martial law.

[12] https://huggingface.co/FacebookAI/xlm-roberta-base

[13] https://huggingface.co/beomi/kcbert-base

[14] When we extended to 10 epochs in an effort to improve performance, this resulted in overfitting to the training data.

[15] https://huggingface.co/LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct

[16] https://huggingface.co/trillionlabs/Trillion-7B-preview

[17] https://huggingface.co/naver-hyperclovax/HyperCLOVAX-SEED-Text-Instruct-1.5B

[18] While sensitive language can affect labeling decisions from a broader perspective, we focused on the factual context

Topics	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
Presidential Office	(1,647 / 118,499)	(643 / 40,031)	(544 / 29,618)	(911 / 67,654)	(949 / 52,487)	(388 / 25,949)	(797 / 48,648)	(1,072 / 57,982)	(1,040 / 71,859)	(738 / 42,551)	(1,119 / 90,794)	(3,425 / 390,653)	(13.2k / 1.03m)
National Assembly / Political Parties	(2,679 / 244,407)	(1,661 / 130,922)	(2,158 / 173,336)	(1,500 / 142,668)	(2,437 / 149,724)	(2,421 / 167,195)	(2,875 / 174,197)	(2,713 / 146,288)	(2,148 / 144,711)	(2,471 / 138,529)	(2,622 / 171,974)	(6,403 / 653,247)	(32.0k / 2.43m)
North Korea	(249 / 15,340)	(70 / 1,977)	(57 / 2,333)	(34 / 1,243)	(43 / 1,738)	(319 / 15,744)	(52 / 1,023)	(106 / 4,340)	(143 / 4,566)	(353 / 16,341)	(192 / 3,943)	(100 / 7,907)	(1.7k / 76.4k)
Administration	(162 / 4,708)	(16 / 359)	(20 / 1,465)	(9 / 146)	(28 / 1,235)	(78 / 3,219)	(23 / 356)	(30 / 1,922)	(35 / 989)	(57 / 2,189)	(23 / 882)	(112 / 7,667)	(0.5k / 25.1k)
National Defense / Foreign Affairs	(130 / 5,126)	(53 / 1,775)	(82 / 2,829)	(64 / 1,495)	(221 / 9,347)	(208 / 8,656)	(96 / 2,773)	(173 / 5,613)	(155 / 7,325)	(110 / 3,917)	(178 / 8,483)	(607 / 35,451)	(2.0k / 92.7k)
General Politics	(7,054 / 672,500)	(5,734 / 489,416)	(7,086 / 572,271)	(4,849 / 495,238)	(961 / 72,292)	(3,272 / 253,428)	(4,427 / 311,983)	(5,001 / 329,824)	(4,362 / 336,104)	(4,398 / 273,979)	(5,203 / 389,716)	(12,655 / 1,422,883)	(65.0k / 5.61m)
Total	(11.9k / 1.06m)	(8.1k / 0.66m)	(9.9k / 0.78m)	(7.3k / 0.70m)	(4.6k / 0.28m)	(6.6k / 0.47m)	(8.2k / 0.53m)	(9.0k / 0.54m)	(7.8k / 0.56m)	(8.1k / 0.47m)	(9.3k / 0.66m)	(23.3k / 2.51m)	(**114k** / **9.28m**)

Table 5: Overview of the total number of political news articles and corresponding comments collected from the Naver in 2024. The values on the left and right represent the number of articles and comments, respectively.

Prior Datasets	Collection Period	*Label Distribution*	Volume
K-Haters	July – August 2021	18.06 : **81.94**	192.1k
KODOLI	October – December 2020	34.62 : **65.38**	38.5k
KoLD	March 2020 – March 2022	4.95 : **95.05**	40.4k
K-MHaS	January 2018 – June 2020	45.65 : **54.35**	109.6k
UnSmile	January 2019 – June 2020	**65.06** : 24.94	18.7k
PoliticalK.O	**January – December 2024**	in Table 1	**9.28m**

Table 6: Comparison of existing and proposed offensive language datasets by key characteristics The label distribution values (left : right) represent the proportion of (offensive : non-offensive) labels.

Prior Datasets	RoBERTa			KcBERT		
	Best (epoch, lr)	*Acc*	*F1*	Best (epoch, lr)	*Acc*	*F1*
K-Haters	5, 1e-5	81.68	75.72	5, 1e-5	83.49	**77.76**
KODOLI	5, 1e-5	89.38	88.07	5, 2e-5	91.22	**90.24**
KoLD	5, 2e-5	80.21	80.20	4, 1e-5	83.84	**83.84**
K-MHaS	4, 1e-5	87.51	87.35	5, 1e-5	89.90	**89.82**
UnSmile	4, 2e-5	84.16	77.86	4, 2e-5	87.57	**83.30**

Table 7: Performance on the re-split test sets of five prior offensive language datasets under optimal conditions.

Topics	Exaone	Trillion	HyperclovaX
Presidential Office	(3.96, 0.41)	(4.75, 0.48)	(4.86, 0.64)
National Assembly / Political Parties	(3.98, 0.31)	(4.73, 0.48)	(4.84, 0.72)
North Korea	(4.02, 0.29)	(4.76, 0.45)	(4.85, 0.68)
Administration	(3.98, 0.39)	(4.73, 0.53)	(4.82, 0.78)
National Defense / Foreign Affairs	(3.97, 0.37)	(4.69, 0.55)	(4.86, 0.67)
General Politics	(4.00, 0.31)	(4.75, 0.45)	(4.86, 0.66)

Table 8: Summary consistency scores (mean, standard deviation) based on self-assessments by each model.

Prior Datasets	Exaone		Trillion		HyperclovaX	
	Acc	*F1*	*Acc*	*F1*	*Acc*	*F1*
K-Haters	82.05	73.66	78.04	73.96	77.59	69.17
KODOLI	76.09	75.54	81.62	80.08	76.09	74.89
KoLD	80.95	80.81	77.96	77.72	74.79	74.75
K-MHaS	66.32	64.32	77.75	77.64	70.32	69.92
UnSmile	85.92	78.98	84.90	80.90	79.46	73.49
Avg	78.26	74.66	80.05	**78.06**	75.65	72.44

Table 9: Performance on the re-split test sets of five prior offensive language datasets under different LLMs.

from news articles published in January and February 2025, ensuring no overlap with the dataset collection period. To mitigate the impact of noise on label predictions (Chen et al., 2023), we corrected the grammar of the sample comments, allowing for a more focused assessment of their offensiveness.

The samples for each topic are provided in Tables 10-15. Expressions like *brainwashed sheeple* and *blind and deaf* in Table 10 reflect extreme language. Similarly, phrases such as *They're so 'brilliant' it's terrifying for our future.* in Table 15 may seem innocuous but convey strong sarcasm targeting a specific group. In contrast, comments that expressed criticism without targeting someone were labeled as non-offensive.

of political discourse, where certain expressions of personal stance are considered acceptable.

B.3 Multi-debate Reasoning Judgment

Model selection This judgment requires five interconnected inferences per comment, and due to the scale of our dataset, applying it across all models would be prohibitively time-consuming. To identify the most suitable model, we evaluated the detection performance of three LLMs using a *Vanilla* prompt on existing offensive language datasets. The results are provided in Table 9.

Trillion achieved an average F1 score of 78.06 and was selected as the primary model for our main analysis. However, its performance remained below that of the fine-tuned PLM under optimal conditions (Table 7). This underscores the limitations of even recent LLMs in zero-shot settings and the need for tailored detection methods.

Sample Comments	Label
계엄은 대통령의 권한이다. 계엄을 내란이란 떠들고 범죄라고 떠드는 놈들은 공산주의 좌파놈들과 공산좌파언론에 속은 개돼지들 틀림없다. (*Martial law is the President's prerogative.* ***Anyone screaming that it's rebellion or calling it a crime is nothing but a brainwashed sheeple, duped by the commie-left and their media puppets.***)	Offensive
윤석열이 1년 넘게 계엄을 입에 담았다는데 용산 종자들이 시각장애우나 청각장애우도 아니고 몰랐을 리가 있겠느냐 전원 출국금지 시키고 내란 공범 여부를 철저하게 조사해야 한다. 저런 것들은 없어도 나라 돌아가는데 아무 지장없다. 시사평론가 장모씨 말로는 요즘은 출근해서 유튜브나 본다고 하더만. (*They say Yoon's been talking about martial law for over a year —* ***do you really think the cronies in Yongsan were blind and deaf the whole time?*** *They all need travel bans and a full investigation into whether they were co-conspirators in treason.* ***Honestly, we'd be better off without these people.*** *According to political commentator Mr. Jang,* ***they don't even work anymore — just sit around watching YouTube all day.***)	Offensive
윤석열 계엄이나 전부 사직으로 막았어야지, 헌법재판관 임명 반대가 더 중요해? (*Yoon Suk-yeol should've stopped this with martial law or by forcing everyone to resign.* *You're telling me opposing a Constitutional Court appointment was more important?*)	Non-offensive
세금이 아깝다. 사표를 수리하고, 내란 동조 관련 혐의는 철저히 수사해야 한다. 비서실장이 계엄을 몰랐을 리 없다. 정진석도 한덕수처럼 윤석열에게 코 꿰어 어쩔 수 없이 지지하는 것처럼 보인다. (*What a waste of taxpayers' money. Their resignations should be accepted, and there must be a thorough investigation into charges of aiding an insurrection.* *There's no way the Chief of Staff didn' t know about the martial law plans.* *Jeong Jin-seok is starting to look just like Han Duck-soo — hooked by Yoon and left with no choice but to support him.*)	Non-offensive

Table 10: Few-shot samples used for the comments under the topic *Presidential Office*.

Sample Comments	Label
그러니까 니가 공화당 따위 대표나 하고 앉아 있는 거야. (***That's exactly why you're stuck as the so-called leader of some wannabe Republican knockoff.***)	Offensive
국힘이 잘해서 지지하는게 아니고 리재명이 싫어서 할 수 없이 견제하는거다. 정치인 다 정신 차려야 된다. (*People aren't supporting the PPP because they're doing a great job* ***— they just can't stand Lee Jae-myung, so they're backing them as a necessary evil.*** *Every politician needs to snap out of it.*)	Offensive
저 인간, 전에 국회에서 난리치려다가 끌려나간 인간이었지 아마? (*Wasn't that guy the one who tried to cause a scene in the National Assembly and ended up getting dragged out?*)	Non-offensive
맞음 국민의 힘 지지율은 국민의힘 지지율이라기 보다 윤석열 대통령 복귀를 위한 지지율임 고로 배신때린 자들은 대권 생각마셈. (*Exactly. The support ratings for the People Power Party aren't really about the party — they're basically a proxy for backing President Yoon's return.* *So anyone who stabbed him in the back shouldn't even dream of running for president.*)	Non-offensive

Table 11: Few-shot samples used for the comments under the topic *National Assembly / Political Parties*.

Sample Comments	Label
공산주의자와는 대화가 안된다. 우리가 핵무기 10개만 있으면 대화하자고 나올 거다. 트럼프는 순진한가나? 바보인가? (***There's no talking with communists.*** *But if we had just ten nukes, they'd be the ones begging to negotiate.* ***What, was Trump just being naive? Or straight-up stupid?***)	Offensive
공산주의 놈들 신났네. 역시 사상과 이념이 공산주의라 트럼프가 만나준다니까? 아주 좋아 죽네. 공산주의 놈들아 니들은 사상이 문제야 사상이. 김정은 한테 욕한번 해봐라. 못하지? 김정은이 그리 좋으냐? 우리 동포들이 지금도 굶어서 죽는다. 이 공산주의 놈들아. (***The commie bastards must be thrilled.*** *Of course they're loving it — Trump's actually agreeing to meet with them.* ***Typical of their twisted ideology.*** ***Hey, why don't you try cursing out Kim Jong-un once? Can't do it, can you? You love that dictator so much?*** *Our fellow Koreans are still starving to death, and* ***you commie scum sit there grinning. It's your rotten ideology — that's the real problem.***)	Offensive
이제 정은이가 콧대 높아져서 남한은 쳐다도 안보겠네 잘됐네. 우리 우파는 이럴수록 더 뭉칩시다. (*Now that little Jung-eun's gotten all high and mighty, he probably won't even bother looking South anymore.* *Good riddance. This is exactly when we conservatives need to stand even stronger together.*)	Non-offensive
헛물켜지마라. 자고로 미국은 결정적일때 우리를 버렸다. 비단 미국만이 아니다. 지금 우크라이나를 봐라. 힘없는 나라는 평화도 없다. 물론 미국에게 고마운 것도 많지만 우리가 이뻐서 도와주는 것만은 아님을 누구나 다 아는 사실이다. 진영 상대를 적으로만 대하지 말고 인정하고 빨리 국력을 키워야한다. (*Don't get your hopes up. History shows the U.S. always bails when it really matters. And it's not just the U.S.* *— look at Ukraine. Weak countries don't get peace. Sure, we owe the U.S. a lot, but let's not kid ourselves — they're not helping us out of love.* *Everyone knows that. Instead of treating political opponents like enemies, we should acknowledge reality and focus on building national strength.*)	Non-offensive

Table 12: Few-shot samples used for the comments under the topic *North Korea*.

Sample Comments	Label
생각이라는게 있긴 한거임? 국무위원들을 그냥 지졸개 정도로 생각했겠지 모지란 것이. (***Do you even have the ability to think? You probably saw the cabinet ministers as nothing more than clueless flunkies — typical of someone that dim.***)	Offensive
계엄자체가 정상이 아닌데 무슨 윤수괴한테 정상적인 것을 기대하냐. 친일도 계몽이라는 놈한테 무슨 정상적인 절차가 있었겠냐. (*Martial law itself is completely abnormal —* ***why would anyone expect anything normal from Yoon the Tyrant?*** ***The guy who called Japanese colonialism 'enlightenment'*** *— you think he cares about following due process?*)	Offensive
어찌되었든 계엄은 잘못된 거라고 생각한다. 윤석열씨는 내려오는게 맞다고 생각한다. (*Regardless of the details, I believe declaring martial law was wrong. Yoon Suk-yeol should step down.*)	Non-offensive
계엄은 대통령과 국방장관이 계획했고, 결심권자는 대통령이고, 국무회의는 형식상의 절차였고, 국무위원은 의결권이 없고, 의견발언권만 있다. 계엄은 합법이고, 설령 위법성이 있다하여도, 대통령은 사소한 법위반으로 기소되지 않는 특권이 있다. 나는 대통령의 비상계엄을 지지한다. (*Martial law was planned by the President and the Defense Minister. The final authority rests solely with the President.* *The cabinet meeting was just a formality — ministers don't have voting power, only the right to express opinions.* *Martial law is legal. And even if some legal issues arise, the President is constitutionally immune from prosecution over minor violations.* *I fully support the President's emergency martial law declaration.*)	Non-offensive

Table 13: Few-shot samples used for the comments under the topic *Administration*.

Sample Comments	Label
간첩이 득실대고 있는데 민주건달들은 무슨 짓을 하고 있는 거냐? (*While spies are crawling all over the place,* ***what the hell are these Demo-thugs even doing?***)	Offensive
1억 받고 中에 블랙요원 신상 넘긴 군무원? 혹시 더듬이당 당원인지 한 번 까봐라. (*A military official who sold out a black agent's identity to China for a hundred million won?* ***Someone check if he's a member of the Stammering Party.***)	Offensive
국가분열행위, 방산비리인데 재산 몰수 후, 사형이 맞지않나? (*This is an act of treason and a massive defense corruption scandal — shouldn't the punishment be asset seizure and the death penalty?*)	Non-offensive
정말 나라 기강이 다 무너지고 있다. 사태가 이 지경인데 민주당은 왜 간첩법 거부하는가? 정말 이해 할 수 없다. (*The entire moral fabric of the nation is collapsing.* *Things are in total chaos, and yet the Democratic Party is rejecting the Anti-Spy Law? I just can't make sense of it.*)	Non-offensive

Table 14: Few-shot samples used for the comments under the topic *National Defense / Foreign Affairs*.

Sample Comments	Label
경기지사면 지사답게 도정에만 좀 제발 좀 신경써라. 뭐 임방한다고 나와서 떠드냐. (*If you're the governor of Gyeonggi Province, then act like one and focus on running the province.* ***Why the hell are you out here livestreaming and ranting like an influencer?***)	Offensive
돈만 찍어내서 뿌려대고 서민 경제 살렸다며, 자신들이 지운 빚은 지들 죽은 뒤의 미래 세대에 물려주는 존나 유능한 진보가 민주당. 너무 유능해서 미래가 걱정된다. (***The oh-so-'competent progressives' of the Democratic Party keep printing money,*** ***throwing it around, and patting themselves on the back for 'saving' the economy*** ***— all while dumping the debt they created on future generations after they're long gone. They're so 'brilliant' it's terrifying for our future.***)	Offensive
보수냐 진보냐 그게 중요한 것이 아니야. 이성적으로 판단하고 세상을 사는 것이 답이야. 이념 탓만 하니 나라 꼴이 이러지. (*It's not about being conservative or progressive — what really matters is thinking rationally and living with common sense.* *This obsession with ideology is exactly why the country's such a mess.*)	Non-offensive
유능한 진보건, 중도 보수건 일 잘하고 국민 삶 윤택해지면 최고다! 청년일자리, 미중 패권전쟁에서 국익 지키는 것, 출산율 저하로 미래 대한민국 존립성 위태로움 극복도 큰과제. 또한 대북 리스크 줄이는 것도 코리아 디스카운트 줄이는 큰 과제일것이다! (*Whether it's competent progressives or moderate conservatives, what matters is results — improving people's lives.* *We need to create jobs for young people, protect our national interests in the U.S.-China power struggle,* *and tackle the existential threat posed by declining birthrates. Reducing the North Korea risk is also key to ending the 'Korea Discount' once and for all.*)	Non-offensive

Table 15: Few-shot samples used for the comments under the topic *General Politics*.

C Further Experimental Results in Three Main Judgments

C.1 Evaluating Judgments on Ground Truth

We conducted human annotation on a subset to evaluate the performance of each judgment against actual ground truth, involving five university graduates who labeled 20 comments across six topics, totaling 120 comments each. Annotators were provided with labeling criteria detailed in Appendix B.2 and example annotations from Tables 10-15. We then aggregated their results via majority voting to derive final labels used as ground truth. The Krippendorff's α was 0.62, indicating moderate agreement among human raters.

The evaluation results of each judgment on a human-annotated subset are presented in Table 16. While PEJ achieved the highest performance in our main analysis, SEJ slightly outperformed it when evaluated against ground truth. This suggests that the offensive language datasets used in SEJ may align closely with the labeling criteria adopted in our study. Although these results are based on only 0.001% of the dataset due to the limited scale of human annotation, we observed a consistent trend of higher scores across all judgments when evaluated using ground trust†. This implies that follow-up analyses should consider the potential overestimation introduced by judgment-based automated evaluations.

Each Judgment		*Acc*	*P*	*R*	*F1*
SEJ	ground truth	65.83	74.02	73.16	65.81
	ground trust†	86.66	80.15	90.10	82.95
PEJ	ground truth	55.00	71.87	65.38	53.96
	ground trust†	97.50	96.87	95.47	96.15
MRJ	ground truth	51.66	71.00	62.82	49.98
	ground trust†	90.83	88.50	82.42	84.95

Table 16: Evaluation of each judgment on a human-annotated subset using both ground truth and ground trust† for all topics.

C.2 MRJ on Other Models

We applied MRJ to the subset corresponding to May 2024 that contains *0.28 million* comments, using Exaone and HyperclovaX with the analysis presented in the Table 17. Both Exaone and Trillion predicted a higher proportion of offensive labels, whereas HyperclovaX showed a lower rate, around 50%. While HyperclovaX exhibited a comparable overlap ratio to other models with SEJ, its ratio with PEJ was notably lower, suggesting that prompt- and debate-level consistency may be more limited in smaller models.

Models	*Label Distribution*	PEJ ⇌ MRJ	MRJ ⇌ SEJ
		Pairwise Label Overlap Ratio	
Exaone	85.39 : 14.60	87.00	69.97
Trillion	81.70 : 18.29	83.91	69.85
HyperclovaX	50.39 : 49.60	64.56	66.73

Table 17: Label distribution and pairwise label overlap ratio using MRJ on a subset across all models and topics.

C.3 Case Study

The agreement examples for each judgment are provided in Table 18. Cases uniquely labeled as non-offensive by SEJ often lacked explicit profanity, but the LLM-based approach appears to have classified them as offensive depending on context. For PEJ, comments conveying implicit hostility toward *hardcore supporters* were still labeled non-offensive. Lastly, non-offensive judgment by MRJ typically involved personal political opinions that, upon an objective perspective, exhibited no clear signs of offensiveness. These cases illustrate the varying interpretations that can arise from different judgment criteria, even for the same comment.

SEJ	PEJ	MRJ
Non-off	*Off*	*Off*
도대체 국민 70%가 찬성하는 김건희 특검은 거부하면서 공정과 상식은 어디에 있는지? (*So they reject a special investigation into Kim Keon-hee that 70% of the public wants, do they still have the nerve to talk about fairness and common sense?*)		
SEJ	PEJ	MRJ
Off	***Non-off***	*Off*
여야 강성들만 보고 정치하면 망하는 지름길이다. (*When you let only the hardcore supporters from both sides shape your politics, you're setting yourself up for failure.*)		
SEJ	PEJ	MRJ
Off	*Off*	***Non-off***
자기 당과 대통령 망하라고 여론전 하는 인간이 여기 있네. (*Funny how he's out there pushing narrative that make it look like he wants his own party and president to fail.*)		

Table 18: A case study uniquely identified as non-offensive by each judgment. *Off* and *Non-off* denote offensive and non-offensive, respectively.

D Prompt Constructions

- The following prompts are used in our **prompt-variants ensemble judgment** (PEJ).

system
Your task is to classify whether the given comment on Korean political news articles is offensive.

user
If $Defn${
Offensive language refers to words or expressions intended to insult, harm, or belittle individuals or groups. Especially in the context of Korean political discussions, this includes:
- Insulting, derogatory, or combative language that may be directed at political figures, parties, or ideologies.
- Discriminatory language based on political affiliation, race, religion, gender, or other personal attributes.
- Sarcastic or harmful humor that may target political beliefs, parties, or individuals in a demeaning way.
- Incitement of violence or hatred against political opponents, minority groups, or societal institutions.
- Misinformation or harmful narratives that spread unverified or damaging political views, especially in a divisive or inflammatory manner.
}
If $Summ${
The following sentences are a summary of the original article on which the comment was posted.

Summary: {*summary of the source article*}
}
If $FewShots${
You may refer to the following examples.

Comment: {*sample comment under the same topic with target comment*}

Label: {*sample pseudo-label under the same topic with target comment*}
...
Comment: {*sample comment under the same topic with target comment*}

Label: {*sample pseudo-label under the same topic with target comment*}
}
If $Vanilla${
Please classify whether the comment is offensive or not. Respond with only "yes" or "no", without any explanations.

Comment: {*target comment*}

Label:
}

- The following prompts are used in the *Summ* prompt for summary generation and evaluation to each model.

Summary generation
system
Your task is to provide consistent summaries of Korean political news articles.

user
Based on the following title and content, please summarize the news article in three sentences.

Title: {*title of the source article*}

Content: {*content of the source article*}

Summary:

Summary evaluation
system
Your task is to evaluate how well the given summary reflects the content of Korean political news articles.

user
Please score the following summary for its consistency with the corresponding article on a scale from 1 to 5, without any explanations. Consistency refers to how much of the information in the summary is actually present in the source article. A score of 5 means that all statements in the summary are fully supported by the article, while a score of 1 means the summary does not reflect the article at all.

Summary: {*summarized article*}

Source Title: {*title of the source article*}

Source Content: {*content of the source article*}

Score:

- The following prompts are used in our **multi-debate reasoning judgment** (MRJ).

Persona alignment for LLM_O (LLM_N)

system
You argue that the given comment from Korean political news is offensive (non-offensive), and provide a clear justification based on linguistic and contextual cues.

user
Please classify whether the comment is offensive or not. Respond with only "yes" or "no", without any explanations.

Comment: {*target comment*}

Label:

assistant
{*yes* (*no*)}

Rationale generation

user
You think the comment is offensive (non-offensive).

The following sentences are a summary of the original article on which the comment was posted.

Summary: {*summary of the source article*}

Based on the context, given a reason why you think the comment is offensive (non-offensive). Please do so in one sentence.

Reason:

Discuss on each side

system
You argue that the given comment from Korean political news is offensive (non-offensive), and provide a clear justification based on linguistic and contextual cues.

user
You have to consider the given comment.

Comment: {*target comment*}

Your argument about this comment:
{$rationale_O$ ($rationale_N$)}

The non-offensive (offensive) side's argument about this comment:
{$rationale_N$ ($rationale_O$)}

Based on your argument, agree or rebut the non-offensive (offensive) sides' argument and explain your reason. Please do so in one sentence.

Stance:

Final judgment

system
Your task is classify whether the given comment on Korean political news articles is offensive. You should consider both sides, offensive and non-offensive, fairly to maintain a balanced perspective.

user
As a judge, assess the debaters' arguments and stances based on the following criteria, "How well they capture the non-offensiveness or offensiveness of the comment". Consider both sides fairly to maintain a balanced perspective and make a broad judgment.

The offensive side's argument: {$rationale_O$}
The non-offensive side's argument: {$rationale_N$}

The offensive side's stance: {$stance_O$}
The non-offensive side's stance: {$stance_N$}

Please classify whether the comment is offensive or not. Respond with only "yes" or "no", without any explanations.

Comment: {*target comment*}

Label: