KOCOH: Korean Context-Dependent Hate Speech Dataset

Eunah Park, Sanghoun Song†
Korea University
Seoul, Republic of Korea
{dmsdk1993, sanghoun}@korea.ac.kr

Abstract

***Warning**: This paper contains hateful and/or offensive contents.*
We introduce the **KO**rean **CO**ntext-dependent **H**ate speech dataset (KOCOH) to evaluate large language models' ability to detect context-dependent hate speech in Korean. KOCOH consists of 3,000 context-comment pairs collected from Korean online communities (Dcinside, FMkorea) with detailed annotations, including labels for hate speech and hate target groups. We assess the context-dependent hate speech detection capabilities of both humans and 11 state-of-the-art large language models, including GPT-5, Claude Sonnet 4, and Gemini 2.5 Flash. Our results show that humans outperform language models, with GPT-5 achieving the highest performance among the evaluated models. While humans demonstrate balanced recall and specificity, language models generally show significantly higher specificity compared to recall. The performance of both humans and models is affected by factors such as *Honam*-related vocabulary and sentiment polarity. This study contributes resources to Korean hate speech research and empirically demonstrates the performance gap between humans and language models. Through both quantitative and qualitative analyses, we explore the similarities and differences between humans and language models, offering insights for future developments in language models and AI ethics research. KOCOH is available at https://github.com/eparkatgithub/KOCOH.

Keywords: Hate Speech Detection, Context-Dependent Hate Speech, Human–Language Model Comparison

1. Introduction

While the performance of language models has improved significantly, ethical awareness and considerations have not advanced at the same pace. Given the potential impact of language models—and artificial intelligence more broadly—on society, ethical considerations remain essential in natural language processing research (Hovy and Spruit, 2016; D'Arcy and Bender, 2023; Tacheva and Ramasubramanian, 2023). In this context, the ability of language models to detect and respond to abusive language must be studied continually from various perspectives. One particularly harmful aspect is hate speech—language that attacks a group or its members based on specific collective characteristics. Hate speech is problematic not only as a manifestation of emotional hostility but also because it can lead to discrimination and violence in the real world, exhibiting high social contagion (Hong, 2019).

In this study, we define hate speech as language that disparages or insults social minorities or vulnerable individuals based on prejudice and discrimination, or that incites discrimination or violence against them (Nockleby, 2000; Nobata et al., 2016; Fortuna and Nunes, 2018). Accordingly, hate speech cannot be identified solely based on the presence of hateful words; understanding requires considering both the semantic meaning of the utterance and its context (Schmidt and Wiegand, 2017; Waseem et al., 2017). For example, "The word 'ni**er' should not be used" contains a hateful term, yet it expresses opposition to hate speech.

Research on hate speech in natural language processing has been ongoing for a long time, resulting in the creation of numerous datasets. However, most previous studies have focused on collecting large volumes of comments from social media or online communities and assessing hateful intent based solely on the content of individual statements (Waseem and Hovy, 2016; Moon et al., 2020; ElSherief et al., 2021; Lee et al., 2022). This approach is effective for statements where the target and intent of the hate are explicitly revealed in the statement itself, such as "I am surprised they reported on this crap who cares about another dead ni**er?" and "300 missiles are cool! Love to see um launched into Tel Aviv! Kill all the gays there!" (Nobata et al., 2016).

However, accurately identifying the hateful intent of an utterance often requires contextual understanding—specifically, analyzing related posts or the comments that precede and follow it. For example, the utterance "Thank you for saying what I was thinking!" is not abusive in itself and even seems positive. Yet, if it is a response to the statement "Because religious people who think they have license from God to do whatever they want are f***ing psychotics," it would be considered hate speech directed at a religious group (Baheti et al., 2021).

† Corresponding author.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4103–4114
11–16 May 2026.

Hate	
Context	Several companies have recently engaged in acts of gender discrimination in their recruitment processes.
Comment	This is the right direction.
Non-Hate	
Context	Several companies have announced welfare benefits policies recently.
Comment	This is the right direction.

Table 1: Examples of hateful and non-hateful comments in various contexts.

Context-dependent hate speech refers to utterances that are perceived as hateful only when interpreted within a given context. As shown in Table 1, **Comment** "This is the right direction." appears in both cases. However, in a context that endorses gender discrimination in recruitment, it conveys hateful intent and is labeled as **Hate**. Conversely, when the same comment appears in a neutral or unrelated context, it is classified as **Non-Hate**.

Detecting context-dependent hate speech is important for two reasons. First, the same utterance can have different meanings depending on the participants in the conversation, their sociocultural backgrounds, and the situational context. Thus, context plays a crucial role in identifying hate speech (Prabhakaran et al., 2020). Second, context-dependent hate speech occurs frequently in real-world communication, yet it remains underexplored, highlighting the need for improvements in hate speech detection systems.

Despite its importance, many studies have assessed the ability of models to understand context and detect hate speech separately, but few have systematically examined both aspects together. As a result, there is a need for a dataset specifically designed to evaluate context-dependent hate speech detection.

Park and Song (2025) introduced the Korean Context-Dependent Hate Speech dataset (KOCOH), the first Korean hate speech dataset specifically designed for context-dependent hate speech. In this study, we present KOCOH v.2, an advanced version of KOCOH. Compared to the initial version (Park and Song, 2025), KOCOH v.2 includes:

- Refined existing instances (modified, supplemented, or deleted when necessary)
- 995 additional instances (from 2,005 to 3,000)
- Expanded data sources
- Enhanced annotation labels (counter speech, profanity, etc.)

The main contributions of this study are summarized as follows:

- We enhance KOCOH, the first context-dependent hate speech dataset in Korean, by modifying and supplementing its existing data and by collecting data from a new platform.
- We empirically demonstrate the performance gap between humans and language models in detecting context-dependent hate speech. Additionally, we discuss potential directions for enhancing model performance through comparative analysis.

2. Related Work

Contextual information is crucial for hate speech detection, as the same utterance can convey different intentions depending on its surrounding context. Early studies acknowledged this significance; Gao and Huang (2017) introduced the first dataset incorporating contextual information. Their research showed that detection performance improves when the model is provided with such information. Consequently, subsequent studies have increasingly concentrated on context-dependent cases where hate speech cannot be identified from isolated utterances alone.

Table 2 lists existing abuse detection datasets in English and Korean that incorporate contextual information. Several studies in English have explicitly addressed context dependence. Baheti et al. (2021) demonstrated that responses agreeing with offensive contexts are highly likely to be offensive themselves. Additionally, Yu et al. (2022) found that annotators' judgments in hate speech identification tasks often changed when context was provided. Sun et al. (2022) and Shin et al. (2024) examined cases where bot responses in chatbot conversations were deemed unsafe or offensive solely due to contextual influence. These studies emphasize that context can significantly alter the interpretation of an utterance from benign to hateful or vice versa.

Conversely, few Korean abuse datasets explicitly account for context-dependence in the same way that English-language datasets do. While Jeong et al. (2022) explored how context influences model performance, their focus was on performance improvement rather than on instances where identical utterances receive different labels based on context. Other datasets have incorporated context for various reasons, such as bias reduction or synthetic data generation, but they do not specifically address context-dependent hate speech, where the same comment can be classified as hateful or non-hateful depending on its context.

Park and Song (2025) addressed this gap by introducing the first Korean dataset specifically designed for context-dependent hate speech detection. We extend their work by expanding the dataset

Authors	Lang	Size	Source	Target	Context-Dependent?
Gao and Huang (2017)	EN	1,528	Fox News	**Hate speech**	✗
Baheti et al. (2021)	EN	2,000	Reddit	Offensive language	✓
Sun et al. (2022)	EN	11,492	Reddit (mainly)	Unsafety	✓
Yu et al. (2022)	EN	6,846	Reddit	**Hate speech**	✓
Shin et al. (2024)	EN	11,271	SimSimi chatbot	Offensive language	✓
Jeong et al. (2022)	**KO**	40,429	Naver News & YouTube	Offensive language	✗
Lee et al. (2023a)	**KO**	139,854	Human & Synthetic data	Sensitive answer	✗
Lee et al. (2023b)	**KO**	34,214	Synthetic data	Bias	✗
Shin et al. (2023)	**KO**	10,000	Twitter & Synthetic data	Offensive language	✗
Kim et al. (2024)	**KO**	22,249	Crowd workers	Violence	✗
Park and Song (2025)	**KO**	2,005	Dcinside	**Hate speech**	✓
Ours	**KO**	3,000	Dcinside & FMkorea	**Hate speech**	✓

Table 2: Comparison of context-aware abuse detection datasets.

and conducting a broader range of experiments, enabling deeper insights and more rigorous analysis.

We present examples where the same comments are interpreted as hateful in one context but non-hateful in another, as well as instances where a single context produces both hateful and non-hateful comments. This contrastive design facilitates a systematic evaluation of the models' ability to understand context and detect hate speech simultaneously. While most prior studies have focused on enhancing model performance by incorporating contextual features, our primary aim is to assess how well models perform on this challenging task and to compare their performance to that of humans. This evaluation-centered approach, combined with our contrastive dataset design, sets our work apart from existing context-aware hate speech datasets in both English and Korean.

3. KOCOH

3.1. Data Collection

We collect data from the popular post boards of Dcinside and FMkorea, two of the most visited Korean online communities.[1]

For Dcinside, we review all posts written between June 2 and June 23, 2024, and for FMkorea, all posts written between January 1 and January 21, 2025. We examine both the post content and comments, collecting context-dependent hate speech comments that meet the definition in Section 1. We also include summaries of the corresponding post content, selecting only unambiguous instances where the context-dependent nature of hate speech is evident. Duplicate comments from the same post are excluded. The collected data are categorized into three types as follows.

As shown in Table 3, we construct the data into three types: **Type 1** pairs a real context (post content) with a hate speech comment from that post. **Type 2** pairs a manually created context with the *same comment* from Type 1, but in this case, the comment is no longer hateful. **Type 3** pairs the *same context* from Type 1 with a different, non-hateful comment from the same post.

To illustrate, consider the examples in Table 3. The Type 1 comment is interpreted as hateful because it portrays Gwangju as if it were foreign territory. This pattern—treating Korean regions as foreign—is a common form of regional hate speech in Korean online communities. In contrast, the same comment becomes non-hateful in Type 2 because Japan is a foreign country, making the passport reference literally accurate rather than discriminatory. Similarly, Type 3 shows that the same context can lead to non-hateful comments: expressing admiration for Gwangju carries no derogatory implications.

This design facilitates two important comparisons: First, comparing Type 1 and Type 2 illustrates how the same comment can be perceived as either hateful or non-hateful depending on its context. Second, comparing Type 1 and Type 3 reveals that a single context can lead to both hateful and non-hateful comments. These configurations enable KOCOH to assess whether language models can accurately determine comment intent based on contextual cues. In total, we collected 3,067 context-comment pairs.

3.2. Annotation Procedure

The 3,067 collected pairs undergo full consensus review by three expert annotators (two researchers and one independent reviewer). Each annotator independently classified every instance, and only unanimous agreements were retained, yielding 3,000 final pairs (97.82% retention rate). Discarded instances (2.18%) involved ambiguous intent or multiple plausible interpretations.

We employ strict consensus instead of traditional

[1] https://waffleboard.io/ranking

Type	Context	Comment	Hate Speech?
Type 1	광주광역시의 맛있는 음식 다섯 가지를 소개했다. Introduced five delicious foods from Gwangju.	먹고 싶은데 여권 들고 가기 귀찮아:: I want to eat them but it's annoying to bring my passport.	✓
Type 2	일본 오사카의 맛있는 음식 다섯 가지를 소개했다. Introduced five delicious foods from Osaka, Japan.	먹고 싶은데 여권 들고 가기 귀찮아:: I want to eat them but it's annoying to bring my passport.	✗
Type 3	광주광역시의 맛있는 음식 다섯 가지를 소개했다. Introduced five delicious foods from Gwangju.	역시 광주다 This is Gwangju for you.	✗

Table 3: Examples of each data type.

inter-annotator agreement metrics because context-dependent hate speech requires specialized cultural knowledge and subjective judgment. This approach prioritizes label quality over coverage, ensuring high-confidence gold standard annotations for evaluating whether models can achieve expert-level performance on clear-cut cases.

Beyond binary hate speech classification, we annotate additional dimensions for each instance:

- **Hate Target Groups**: Since a single comment can express hate toward multiple targets, these categories are not mutually exclusive. We identify five hate target groups present in our dataset:
  - **Gender**: women and sexual minorities,
  - **Disability**: people with disabilities,
  - **Race/Nationality**: specific races, countries, or ethnic groups,
  - **Region (Korea)**: regional groups within Korea such as *Honam* (a region located in southwestern Korea, which can be considered a social minority within Korean society),
  - **Age**: particular age groups such as the elderly or children.
- **Counter Speech**: Counter speech refers to responses that oppose and reject hate speech, aiming to prevent its spread and reduce its impact (Gelber, 2002).
- **Profanity**: This is a supplementary label for broader abusive language research.

3.3. Annotation Results

Table 4 presents the number of instances for each data type, and Table 5 shows the distribution of data across hate target groups. The dataset comprises

Source	Type	Count (%)
Total	Type 1	764 (25.47)
	Type 2	764 (25.47)
	Type 3	1,472 (49.07)
		3,000 (100.00)
Dcinside	Type 1	529 (17.63)
	Type 2	529 (17.63)
	Type 3	924 (30.80)
		1,982 (66.07)
FMkorea	Type 1	235 (7.83)
	Type 2	235 (7.83)
	Type 3	548 (18.27)
		1,018 (33.93)

Table 4: Number of instances by data type.

886 hate speech instances (29.53%) and 2,114 non-hate instances (70.47%).

Notably, the dataset contains very few instances related to **Disability**. Hate speech directed at **Disability** is explicitly present in both the Dcinside and FMkorea communities, which makes it challenging to collect examples of context-dependent hate speech in this category. Similarly, the scarcity of **Region**-targeted data in FMkorea stems from the fact that regional hate speech on that platform often appears explicitly.

Additionally, we report statistics for supplementary annotations: counter speech (57 instances) and profanity (277 instances).

4. Experimental Settings

4.1. Language Model

Models Eleven state-of-the-art large language models were used in our experiments. These include globally recognized models such as GPT-5, Claude Sonnet 4, and Gemini 2.5 Flash, along with

Source	Target Group	Count (%)
Total	Gender	370 (48.43)
	Disability	23 (3.01)
	Race/Nation	199 (26.05)
	Region	215 (28.14)
	Age	79 (10.34)
Dcinside	Gender	193 (25.26)
	Disability	14 (1.83)
	Race/Nation	131 (17.15)
	Region	203 (26.57)
	Age	64 (8.38)
FMkorea	Gender	177 (23.17)
	Disability	9 (1.18)
	Race/Nation	68 (8.90)
	Region	12 (1.57)
	Age	15 (1.96)

Table 5: Number of instances by hate target group. As described in Section 3.2, hate target groups are not exclusively annotated, so the total percentage exceeds 100%.

the Korean-developed models A.X 4.0, EXAONE 4.0, Solar Pro 2, Tri, Kanana 1.5, DNA 2.0, HyperCLOVA X SEED Think (CLOVA), and Mi:dm 2.0.

For models that support reasoning mode, we typically use the non-reasoning (default) mode. However, to examine the impact of reasoning mode on performance, we also compare the results for both reasoning and non-reasoning modes for two representative models: GPT-5 and CLOVA. As a result, we present a total of 13 model-specific results. Table 6 provides descriptions of each model.

GPU and Hyperparameters We perform inference on open-source models using a single NVIDIA A100 80GB PCIe GPU with 8-bit quantization to optimize memory efficiency. For GPT-5, the verbosity parameter is set to low. In reasoning mode, the reasoning effort parameter is configured as medium, while in non-reasoning mode, it is set to minimal. For all other models, the temperature is adjusted to its minimum value allowed by each model (0-0.1) and all remaining hyperparameters are maintained at their default settings.

Prompt We begin by providing the model with an explanation of the task objective, followed by a definition of hate speech. Next, we conduct a binary classification of hate speech based on the given context, using a zero-shot setting. Each model is instructed to directly output a value of 0 or 1, representing non-hate and hate, respectively. All content is written in Korean.

4.2. Human

Experimental Design Following Ruis et al. (2023) and Divya Venkatesh et al. (2024), we divide the entire set of 3,000 instances into 60 subsets of 50 items each, assigning one subset to each participant. During sampling, we ensure minimal post duplication within the subsets; when duplication is unavoidable, we verify that the corresponding comment content does not overlap.

Quality Control via Pre-test Items Each subset includes 10 additional pre-test items with known correct labels, which are randomly inserted among the 50 evaluation items to assess participant reliability. These pre-test items were collected and annotated separately using the same three-expert unanimous consensus procedure described in Section 3.2, ensuring they meet the same quality standards as the main dataset. The pre-test items serve as attention checks; only participants achieving $\geq$90% accuracy on these items are included in the final analysis.

Annotation Task and Participants Participants were instructed to classify each context-comment pair as either hate speech or non-hate speech, using the same definition provided to the language models (Section 4.1). A total of 62 people participated, but two were excluded based on their pre-test scores, resulting in 60 valid participants.

Performance Calculation We evaluate human performance through the following procedure.

1. Each of the 3,000 instances is evaluated by exactly one participant (non-overlapping assignment across 60 subsets).

2. For each instance, we compare the participant's binary judgment (hate/non-hate) against the gold label.

3. Aggregate metrics (Table 7) are computed by pooling all 3,000 individual judgments.

4.3. Metric

Due to the imbalanced class distribution in KOCOH, we use the F1 score as the primary evaluation metric. To determine if model and human performance exceeds random predictions, we also calculate the zero-rate baseline, which is defined as the proportion of the most frequently occurring class. Accuracy is then compared to this baseline, and cases where accuracy falls below the zero-rate are excluded from further analysis.

Models	Versions	Size	Type	Reasoning?
GPT-5	gpt-5	-	API	✓/✗
Claude Sonnet 4	claude-sonnet-4-20250514	-	API	✗
Gemini 2.5 Flash	gemini-2.5-flash	-	API	✗
A.X 4.0	skt/A.X-4.0	71.9B	API	✗
EXAONE 4.0	LGAI-EXAONE/EXAONE-4.0.1-32B	32.0B	Open-source	✗
Solar Pro 2	solar-pro2	31.0B	API	✗
Tri	trillionlabs/Tri-21B	20.7B	Open-source	✗
Kanana 1.5	kakaocorp/kanana-1.5-15.7b-a3b-instruct	15.7B	Open-source	✗
DNA 2.0	dnotitia/DNA-2.0-14B	14.8B	Open-source	✗
CLOVA	naver-hyperclovax/HyperCLOVAX-SEED-Think-14B	14.7B	Open-source	✓/✗
Mi:dm 2.0	K-intelligence/Midm-2.0-Base-Instruct	11.5B	Open-source	✗

Table 6: Large language models used in the experiments.

5. Experimental Results

5.1. Preliminary Experiment

To investigate the overall ability of language models to detect hate speech, we first conduct a preliminary experiment using a straightforward test set. We select 100 context-independent and unambiguous instances of hate speech from KOLD (Jeong et al., 2022), a Korean hate speech dataset that includes contextual information. The models perform binary classification using the same prompt as in the main experiment.

As a result, the average model accuracy is 97.62% (standard deviation: 3.84), indicating that the language models are highly effective at detecting hate speech. However, a critical question remains: does this high performance continue when the task requires context-dependent comprehension?

5.2. Main Experiment

Table 7 summarizes the main experimental results. Overall, humans outperform language models in detecting context-dependent hate speech. The human F1 score is 6.21 percentage points higher than that of the best-performing model. A stratified bootstrap test (10,000 iterations) with Holm's multiple comparison correction shows that the human F1 score is statistically significantly higher than all models ($p < 0.0001$ and Cohen's d > 3 for all pairwise comparisons).

Notably, although humans have lower precision than the language models' average of 67.93%, their recall is more than double the models' average of 41.57%. In other words, language models tend to minimize false positives, while humans prioritize reducing false negatives.

The average accuracy of the language models is 78.71%, which is 4.18 percentage points higher than the Zero-Rate baseline (i.e., the proportion of the majority class). However, Tri and Kanana 1.5 fall below this baseline.

Although overall accuracy remains relatively high and consistent across models, a clear asymmetry between recall and specificity emerges. For instance, while the lowest specificity among the models is 86.27%, the highest recall is only 51.96%. This indicates that the models recognize non-hateful intent significantly more reliably than hateful intent. This contrast highlights that, although models can effectively detect non-hate speech based on context, they struggle to accurately identify hate speech. Compared to the preliminary experiment (average accuracy: 97.62%), this finding demonstrates that detecting context-dependent hate speech is a considerably more challenging task for language models.

In contrast, human performance is both higher and more balanced: recall = 84.29%, specificity = 84.84%. This balance indicates that, when judging intent based on context, humans do not asymmetrically rely on whether a comment is hateful. In conclusion, humans not only surpass language models in identifying hate speech within context but also demonstrate greater consistency across intent types. This suggests that current language models have not yet fully achieved human-level context understanding in this task.

6. Human vs. Language Model

In this section, we analyze the similarities and differences between humans and language models in detecting hate speech based on context. As mentioned earlier, the results of Tri and Kanana 1.5 are excluded from the following analysis because they do not meet the baseline requirements.

6.1. Factor 1: *Honam*-Related Vocabulary

(1) a. Context: 전주 제지공장에서 일하던 19세 청년이 쓰러진 채 발견되었고 병원으로 이송된 후 사망하였다.
'A 19-year-old paper mill worker in Jeonju

Models	F1 Score	Precision	Recall	Specificity	Accuracy
Zero-Rate	-	-	-	-	74.53
Human	**73.73**	65.51	**84.29**	84.84	84.70
GPT-5 (R)	67.52	**96.36**	51.96	**99.33**	**87.27**
GPT-5 (N)	62.69	85.52	49.48	97.14	85.00
Claude Sonnet 4	60.67	84.74	47.25	97.09	84.40
Gemini 2.5 Flash	53.12	92.23	37.30	98.93	83.23
A.X 4.0	42.16	83.98	28.14	98.17	80.33
EXAONE 4.0	49.35	54.60	45.03	87.21	76.47
Solar Pro 2	49.47	65.03	39.92	92.67	79.23
Tri	50.23	41.73	63.09	69.90	68.17
Kanana 1.5	25.20	28.84	22.38	81.13	66.17
DNA 2.0	39.21	53.26	31.02	90.70	75.50
CLOVA (R)	57.75	77.36	46.07	95.39	82.83
CLOVA (N)	48.86	68.97	37.83	94.19	79.83
Mi:dm 2.0	45.23	50.48	40.97	86.27	74.73

Table 7: Experimental results. 'R' and 'N' denote reasoning mode and non-reasoning mode, respectively. Values below the zero-rate baseline are underlined. Recall corresponds to accuracy for hate-speech instances, and specificity corresponds to accuracy for non-hate speech instances.

collapsed and was found unconscious, then died after being transported to the hospital.'

b. Comment: 전라도 일상이네
'Typical Jeolla Province'

Unlike other target groups, hate speech targeting **Region** often includes explicit lexical indicators, such as *Honam*-related vocabulary (e.g. 전라도 ('Jeolla Province') and 광주 ('Gwangju') as in (1b). Of the 215 instances of hate speech directed at **Region**, 96 cases (44.65%) contain *Honam*-related vocabulary.

This suggests that the relatively high accuracy in detecting hate speech directed at **Region** may not stem from genuine contextual understanding but rather from a keyword-based judgment. Indeed, a chi-square test of independence between the inclusion of *Honam*-related vocabulary and positive judgment reveals a statistically significant relationship for both humans and all language models ($p < 0.001$, $df = 1$).

This indicates that *Honam*-related vocabulary, while not inherently hateful, serves as a significant lexical cue for identifying hate speech among both humans and models. This tendency is likely related to the fact that the *Honam* region is a frequent target of attacks in Korea (Yang, 2022). Similar phenomena have been reported in prior research, where identity terms referring to social minorities (e.g., queer, woman) can intentionally trigger false positives in hate speech detection (Park et al., 2018; Röttger et al., 2021; Tint, 2025).

Polarity	Count (%)
Positive	349 (45.68)
Negative	162 (21.20)
Neutral	253 (33.12)

Table 8: Sentiment polarity distribution in Type 1 data.

6.2. Factor 2: Sentiment Polarity

Table 8 presents the results of the sentiment polarity classification for Type 1 data, using a classifier trained on KOTE(Jeon et al., 2024), a Korean text emotion dataset. According to the findings of Baheti et al. (2021), comments that respond positively to a hateful context are highly likely to be hate speech. While humans can effectively recognize this correlation, language models often classify comments as non-hate speech if the comment's sentiment is positive, regardless of the context. This phenomenon contributes to the lower recall of language models compared to humans.

(2) a. Context: 일본 만화가 무라타 유스케가 "페미는 인류의 적"이라며 페미니스트가 예술을 망친다고 비난했다.
'Japanese manga artist Yusuke Murata criticized feminists, stating "Feminists are the enemies of humanity" and that they destroy art.'

b. Comment: 참 예술가네
'Such an artist'

Example (2) illustrates a case that humans correctly identified but most language models misclassified. The comment in (2b) expresses praise and

Models	Positive	Negative	Neutral
Human	83.38	84.57	**85.38**
GPT-5 (R)	**56.73**	51.23	45.85
GPT-5 (N)	48.71	**52.47**	48.62
Claude Sonnet 4	39.54	**60.49**	49.41
Gemini 2.5 Flash	29.23	**50.62**	39.92
A.X 4.0	22.35	**38.27**	29.64
EXAONE 4.0	44.70	**45.68**	45.06
Solar Pro 2	29.51	**53.70**	45.45
DNA 2.0	22.06	**45.68**	33.99
CLOVA (R)	**48.42**	43.21	44.66
CLOVA (N)	36.10	**44.44**	35.97
Mi:dm 2.0	41.83	**42.59**	38.74

Table 9: Accuracy of Type 1 data by sentiment polarity.

Models	Included	Not	|Diff|
Human	83.90	84.37	0.47
GPT-5 (R)	25.42	56.81	31.39
GPT-5 (N)	22.88	54.33	31.45
Claude Sonnet 4	28.81	50.62	21.81
Gemini 2.5 Flash	41.53	36.53	5.00
A.X 4.0	22.03	29.26	7.23
EXAONE 4.0	34.75	46.90	12.15
Solar Pro 2	30.51	41.64	11.13
DNA 2.0	23.73	32.35	8.62
CLOVA (R)	31.36	48.76	17.40
CLOVA (N)	25.42	40.09	14.67
Mi:dm 2.0	26.27	43.65	17.38

Table 10: Accuracy of Type 1 data by slang inclusion.

approval of a hateful generalization about feminists. However, most language models fail to associate the positive sentiment with the hateful context, resulting in its misclassification as non-hate speech.

Table 9 presents the accuracy of Type 1 data categorized by sentiment polarity. As shown, all language models, except those in reasoning mode, exhibit lower accuracy for positive-sentiment comments compared to negative ones. In contrast, humans do not show a significant difference in accuracy based on sentiment polarity. A chi-square test of independence confirms that the relationship between sentiment polarity and accuracy is not significant for humans ($p = 0.798$). Among the 9 language models with lower accuracy for positive-polarity instances, 5 models (Claude Sonnet 4, Gemini 2.5 Flash, A.X 4.0, Solar Pro 2, DNA 2.0) demonstrate a significant relationship ($p < 0.001$).

The results suggest that language models may depend more on superficial sentiment cues than humans when detecting hate speech. Consequently, comments expressing a positive sentiment towards hateful actions or situations are more likely to be misclassified as non-hate speech.

6.3. Factor 3: Slang

Dcinside serves as a major source of internet memes and humor that often spread to other communities. However, these neologisms can be difficult to interpret for those unfamiliar with the platform's culture, as their meanings are highly context-dependent and they often function as slang. This is a result of the closed and tribalistic nature of the community (Park, 2022). Consequently, such slang is often created with hateful intent.

To understand slang, one must grasp the context and possess cultural knowledge, including the meanings of the slang and commonly used metaphors. Therefore, humans are likely to have an advantage in classifying data that contains slang.

(3) a. Context: 번역가가 일본 소설을 번역하며 유모차 대신 유아차라는 표현을 써 논쟁이 벌어졌다.
'Controversy arose when a translator used the term *yuacha* (literally "infant vehicle") instead of the more common *yumocha* ("baby carriage") in translating a Japanese novel.'[2]

b. Comment: P
'Blood'[3]

As illustrated in Example (3), there are several instances involving slang where humans successfully identify hate speech, but all language models fail to do so.

In Type 1 data, 118 cases contain slang. As shown in Table 10, all language models except Gemini 2.5 Flash exhibit lower accuracy in the presence of slang, with a particularly pronounced performance difference compared to humans. A chi-square test of independence indicates that for humans, the relationship between slang inclusion and accuracy is not statistically significant ($p = 0.898$). In contrast, of the 10 language models that demonstrate lower accuracy with slang-included data, 8 models (GPT-5 (R), GPT-5 (N), Claude Sonnet 4, EXAONE 4.0, Solar Pro 2, CLOVA (R), CLOVA (N), Mi:dm 2.0) show a significant correlation ($p < 0.05$).

These results suggest that language models have a limited ability to interpret community-specific slang. In contrast, humans seem more adept at using cultural and contextual cues to infer the underlying intent of such utterances.

[2]유모차 (*yumocha*; 'baby carriage') incorporates the Japanese-derived element 유모 (*yumo*; 'wet nurse'), which some feminists have criticized for embedding gendered caregiver assumptions. As a gender-neutral alternative, 유아차 (*yuacha*; 'infant vehicle') has been proposed.

[3]The letter 'P' is pronounced identically to the Korean word 피 (*phi*; 'blood'). This association with women's menstruation is used to disparage women or feminists.

Models	Profanity (↓)	Counter Speech (↓)
Human	13.73	15.79
GPT-5 (R)	0.00	0.00
GPT-5 (N)	2.15	3.51
Claude Sonnet 4	4.29	12.28
Gemini 2.5 Flash	0.43	0.00
A.X 4.0	10.30	8.77
EXAONE 4.0	39.48	54.39
Solar Pro 2	19.31	36.84
DNA 2.0	24.89	36.84
CLOVA (R)	11.59	10.53
CLOVA (N)	19.31	31.58
Mi:dm 2.0	33.05	49.12

Table 11: False positive rates (%) for non-hate instances containing profanity and counter speech.

6.4. Factor 4: Profanity and Counter Speech

The presence of profane or hateful words does not automatically indicate hate speech without considering the context. Among the 277 instances of profanity in KOCOH, 84.12% (233 instances) are labeled as non-hate speech. Similarly, counter speech may include offensive terms while still expressing an anti-hate intent.

Table 11 presents the false positive rates for non-hateful instances that contain profanity and counter speech. The four best-performing models are GPT-5 (R/N), Claude Sonnet 4, and Gemini 2.5 Flash, which exhibit substantially lower error rates compared to models developed in Korea.

Example 4 illustrates a case where humans and the best-performing models correctly identify counter speech containing profanity, while most Korean models misclassify it as hate speech.

(4) a. Context: 초, 중, 고등학교에서 다문화 가정 자녀에 대한 학교 폭력이 무분별하게 이루어지고 있다.
'School violence against children from multicultural families is occurring indiscriminately in elementary, middle, and high schools.'

b. Comment: 퍽킹 레이시스트
'F***ing racists'

Humans and best models accurately identify this comment as criticism of racism rather than hate speech. In contrast, most Korean models misclassify it, interpreting "racist" as a racial slur. Similar patterns emerge with terms like "misogyny" and "neo-Nazi".

Misclassifying counter speech as hate speech is particularly problematic. Counter speech is recognized as an effective means of protecting minorities while preserving free expression (Yu et al., 2022). False positives in counter speech effectively deprive minorities of their means of resistance. The greater prevalence of this issue in Korean models may reflect the scarcity of research and training data on Korean counter speech. Future model development must seriously address this aspect of hate speech detection.

7. Conclusion

In this paper, we release KOCOH v.2, an updated version of the first Korean context-dependent hate speech dataset. Using KOCOH, we systematically evaluate the performance of language models in detecting context-dependent hate speech in Korean and compare their results with those of human participants. Overall, human performance is not only higher but also more balanced across classes compared to that of language models.

The comparison of humans and language models can be summarized as follows: both exhibit a strong association between the inclusion of *Honam*-related vocabulary and positive classifications. However, language models are more susceptible to bias based on sentiment polarity, resulting in lower recall for hateful intent in positive comments. Additionally, they are less accurate with data containing slang, which often requires an understanding of specific cultural or community contexts. Furthermore, many models, particularly those developed in Korea, show substantially higher false positive rates for profanity and counter speech, often misclassifying legitimate resistance from minorities as hate speech.

Our findings reinforce previous studies that emphasize the crucial role of context in hate speech detection and empirically demonstrate that current language models still fall short of human-level context-dependent understanding. We anticipate that KOCOH will serve as a valuable reference point for future research in natural language processing and AI ethics. We hope this work will stimulate the broader development of context-dependent hate speech datasets and detection frameworks in Korean NLP and beyond.

8. Limitations

This study has several limitations. First, the data sources are restricted to online community platforms. Future data collection from social media platforms such as Twitter, news sites, and YouTube will be necessary to obtain a broader and more representative dataset.

Second, the contexts in KOCOH are collected as summaries rather than as the full content of posts. This design choice is based on two practical considerations: the limited context length of current language models and the prevalence of image-based

content in original posts. In the future, it may be beneficial to incorporate visual information or include the full text of posts to enhance contextual richness.

Third, each language model experiment was conducted only once per model, and each human participant was assigned to a single subset of data. These constraints arise from limited resources. Future research could improve the reliability of the results by increasing the number of human participants and conducting repeated model experiments with varied prompting strategies.

9. Ethical Statement

All human-subject experiments conducted in this study were approved by the Korea University Institutional Review Board (KUIRB-2025-0357-020). Participants were informed in advance that they could withdraw from the study at any time if they felt physical or mental discomfort. Adequate compensation was provided after the experiment.

To minimize fatigue among researchers and reviewers during data collection and annotation, sufficient rest periods were implemented. Additionally, measures were taken to prevent potential psychological distress when reviewing or annotating hate speech content.

10. Bibliographical References

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexandra D'Arcy and Emily M Bender. 2023. Ethics in linguistics. *Annual Review of Linguistics*, 9(1):49–69.

Jeevithashree Divya Venkatesh, Aparajita Jaiswal, and Gaurav Nanda. 2024. Comparing human text classification performance and explainability with large language and machine learning models using eye-tracking. *Scientific reports*, 14(1):14295.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Katharine Gelber. 2002. *Speaking Back: The Free Speech Versus Hate Speech Debate*. Discourse approaches to politics, society, and culture. J. Benjamins Publishing Company.

Sung Soo Hong. 2019. Law and policy on hate. *Journal of Law*, 30(2):191–228.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Duyoung Jeon, Junho Lee, and Cheongtag Kim. 2024. User guide for KOTE: Korean online thatgul emotions dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17254–17270, Torino, Italia. ELRA and ICCL.

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.

Minju Kim, Heuiyeen Yeen, and Myoung-Wan Koo. 2024. Towards context-based violence detection: A Korean crime dialogue dataset. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 603–623, St. Julian's, Malta. Association for Computational Linguistics.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Meeyoung Cha, Yejin Choi, Byoungpil Kim, Gunhee Kim, Eun-Ju Lee, Yong Lim, Alice Oh, Sangchul Park, and Jung-Woo Ha. 2023a. SQuARe: A large-scale dataset of sensitive questions and acceptable responses created through human-machine collaboration. In *Proceedings*

of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6692–6712, Toronto, Canada. Association for Computational Linguistics.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. 2023b. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada. Association for Computational Linguistics.

Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han. 2022. K-MHaS: A multi-label hate speech detection dataset in Korean online news comment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3530–3538, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

John T Nockleby. 2000. Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.

Eunah Park and Sanghoun Song. 2025. Kocoh: A dataset for detecting context-dependent hate speech. *Korean Linguistics*, 106:251–277.

In-seoung Park. 2022. Tribalism in the internet community reading memes and neologisms. *Journal of Popular Narrative*, 28(2):59–93.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Zeerak Waseem, Seyi Akiwowo, and Bertie Vidgen. 2020. Online abuse and human rights: WOAH satellite session at RightsCon 2020. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 1–6, Online. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: fine-tuning strategy matters for implicature resolution by llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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.

Jisu Shin, Hoyun Song, Huije Lee, Fitsum Gaim, and Jong Park. 2023. Generation of Korean offensive language by leveraging large language models via prompt design. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 960–979, Nusa Dua, Bali. Association for Computational Linguistics.

Mingi Shin, Hyojin Chin, Hyeonho Song, Yubin Choi, Junghoi Choi, and Meeyoung Cha. 2024. Context-aware offensive language detection in human-chatbot conversations. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 270–277.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. On the safety of conversational models: Taxonomy, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3906–3923, Dublin, Ireland. Association for Computational Linguistics.

Jasmina Tacheva and Srividya Ramasubramanian. 2023. Ai empire: Unraveling the interlocking sys-

tems of oppression in generative ai's global order. *Big Data & Society*, 10(2):20539517231219241.

Joshua Tint. 2025. Guardrails, not guidance: Understanding responses to LGBTQ+ language in large language models. In *Proceedings of the Queer in AI Workshop*, pages 6–16, Hybrid format (in-person and virtual). Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Hye-seung Yang. 2022. Others who are called 'hongeo': A text mining analysis of hate comments towards jeolla province in the context of naver crime news. *Locality & Communication*, 26(2):36–70.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate speech and counter speech detection: Conversational context does matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.